

ISBN: 978-1-7281-1002-8

Proceedings of 2019 IEEE East-West Design & Test Symposium (EWDTS)



Batumi, Georgia, September 13 – 16, 2019

Proceedings of 2019 IEEE East-West Design & Test Symposium (EWDTS)

Copyright © 2019 by the Institute of Electrical and Electronic Engineers, Inc

All Rights Reserved

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright and Reprint Permission: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For reprint or republication permission, email to IEEE Copyrights Manager at pubs-permissions@ieee.org. All rights reserved. Copyright ©2019 by IEEE.

IEEE Catalog Numbers:

USB: CFP19DTW-USB

ISBN: 978-1-7281-1002-8

IEEE Conference Operations

445 Hoes Lane

Piscataway, NJ 08854 USA

Fax: +1 732 981 1769

Email: conference-ops@ieee.org

IEEE: Advancing Technology for Humanity



IEEE



TTTC: Test Technology Technical Council

TTTC IN GENERAL

PURPOSE: The Test Technology Technical Council is a volunteer professional organization sponsored by the IEEE Computer Society and in-cooperation with IEEE CEDA and IEEE Philadelphia Section. The goals of TTTC are to contribute to members' professional development and advancement and to help them solve engineering problems in electronic test, and help advance the state-of-the art. In particular, TTTC aims at facilitating the knowledge flow in an integrated manner, to ensure overall quality in terms of technical excellence, fairness, openness, and equal opportunities.

MEMBERSHIP: Membership is open to individuals interested in test at a professional level.

DUES: There are NO dues for TTTC membership and no parent-organization membership requirements.

BENEFITS: The TTTC members benefit from personal association with other test professionals. They may have the opportunity to be involved on a wide range of committees. They receive appropriate and updated information and announcements. There are substantial reductions in fees for TTTC-sponsored meetings and tutorials for members of IEEE and/or IEEE Computer Society.

TTTC ACTIVITIES

TECHNICAL MEETINGS: To spread technical knowledge and advance the state-of-the art, TTTC sponsors many well-known conferences and symposia and holds numerous regional and topical workshops worldwide.

STANDARDS: TTTC initiates, nurtures and encourages new test standards. TTTC-initiated Working Groups have produced numerous IEEE standards, including the 1149 series used throughout the industry.

TECHNICAL ACTIVITIES: TTTC sponsors a number of Technical Activity Committees (TACs) that address emerging test technology topics and guide a wide range of activities.

TUTORIALS and EDUCATION: TTTC sponsors a comprehensive *Test Technology Educational Program (TTEP)*. This program provides opportunities for design and test professionals to update and expand their knowledge base in test technology, and to earn official accreditation from IEEE TTTC, upon the completion of four full day tutorials proposed by TTEP.

TTTC CONTACT

TTTC On-Line: The TTTC Web Site at <http://tab.computer.org/tttc> offers samples of the TTTC Newsletter, information about technical activities, conferences, workshops and standards, and links to the Web pages of a number of TTTC-sponsored technical meetings.

Becoming a Member: Becoming a TTTC member is extremely simple. You may either contact by phone or e-mail the TTTC office, or fill out and submit a TTTC application form, or visit the membership section of the TTTC web site.

TTTC OFFICE: 1 Marsh Elder Lane, Savannah, GA 31411, USA
Phone: +1-540-937-5066 Fax: +1-540-937-7848 E-mail: tttc@computer.org

TTTC Officers for 2018

Chair

1st Vice Chair

2nd Vice Chair

President of Board

Past Chair

Senior Past Chair

IEEE Design & Test EIC

ITC General Chair

Test Week Coordinator

Secretary

Vice Secretary

Finance Chair

Finance Vice-Chair

Chen-Huan CHIANG Intel - USA

Giorgio Di Natale LIRMM – France

Xiaowei LI Chinese Academy of Science - China

Yervant ZORIAN Synopsys Inc. - USA

Michael NICOLAIDIS TIMA Laboratory - France

Adit D. SINGH Auburn Univ. - USA

Jörg HENKEL Karlsruhe Institute of Technology - German

Li-C WANG UC Santa Barbara - USA

Yervant ZORIAN Synopsys Inc. - USA

André IVANOV U. of British Columbia - Canada

Adam OSSEIRAN Edith Cowan U. – Australia

Peilin SONG IBM - USA

Kenneth D. Mandl - USA

chen-huan.chiang@intel.com

giorgio.dinatale@lirmm.fr

lxw@ict.ac.cn

yervant.zorian@synopsys.com

michael.nicolaidis@imag.fr

adsingh@eng.auburn.edu

henkel@kit.edu

licwang@ece.ucsb.edu

Yervant.Zorian@synopsys.com

ivanov@ece.ubc.ca

a.osseiran@ecu.edu.au

psong@us.ibm.com

mandlken@aol.com

Group Chairs

Technical Meetings

Technical Activities

Tutorials & Education

Standards

Communications

Standing Committee

Industry Advisory Board

Electronic Media

Asia & Pacific

Europe

Latin America

North America

Middle East & Africa

Stefano Di CARLO Politecnico di Torino – Italy

Claude THIBEAULT Ecole Tech. Supérieure – Canada

Paolo BERNARDI Politecnico di Torino – Italy

Adam CRON Synopsys Inc. - USA

Michele PORTOLAN TIMA Laboratory - France

Cecilia METRA U. of Bologna - Italy

Yervant ZORIAN Synopsys Inc. - USA

Giorgio Di Natale LIRMM – France

Kuen-Jong LEE NCKU – Taiwan, R.O.C.

Alberto BOSIO École Centrale LYON – France

Victor Hugo CHAMPAC Inst. Natl. de Astrofísica - Mexico

André IVANOV U. of British Columbia - Canada

Rafic MAKKI GLOBALFOUNDRIES – UAE

stefano.dicarlo@polito.it

thibeault@ele.etsmtl.ca

paolo.bernardi@polito.it

Adam.Cron@synopsys.com

Michele.portolan@imag.fr

cmetra@deis.unibo.it

Yervant.Zorian@synopsys.com

giorgio.dinatale@lirmm.fr

kjlee@mail.ncku.edu.tw

alberto.bosio@ec-lyon.fr

champac@inaoep.mx

ivanov@ece.ubc.ca

raficzein.makki@globalfoundries.com

Technical Activity Committees

Automotive Reliability & Test

Board Testing

Defect Tolerance

Economics of Test

FPGA Testing

Infrastructure IP

Memory Testing

MEMs Testing

Mixed-Signal Testing

Nanometer Testing

Nanotechnology Test

Network-On-Chip Test

On-Line Testing

RF Testing

Silicon Debug and Diagnosis

System Test

3D Testing

Test Compression

Test & Verification

Test Education

Test Synthesis

Thermal Testing

Yervant ZORIAN Synopsys Inc. - USA

Bill EKLOW - USA

Vincenzo PIURI Politecnico di Milano - Italy

Magdy S. ABADIR - USA

Michel RENOVELL LIRMM - France

Yervant ZORIAN Synopsys Inc. - USA

Yervant ZORIAN Synopsys Inc. - USA

Ronald D. BLANTON Carnegie-Mellon U. - USA

Bernard COURTOIS - France

Bozena KAMINSKA IMS Pultronics, Inc. - USA

Jaume SEGURA U. of the Balearic Islands - Spain

Fabrizio LOMBARDI Northeastern U. - USA

Erik Jan MARINISSEN IMEC - Belgium

Michael NICOLAIDIS TIMA - France

Iboun Taimiya SYLLA Texas Instruments - USA

Michael RICCHETTI Synopsys, Inc. - USA

Ian HARRIS UC Irvine - USA

Yervant ZORIAN Synopsys – USA

Rohit KAPUR Synopsys, Inc. - USA

Magdy S. ABADIR - USA

Sule OZEV ASU - USA

Scott DAVIDSON Oracle - USA

Bernard COURTOIS - France

Yervant.Zorian@synopsys.com

beklow56@gmail.com

piuri@elet.polimi.it

magdy.abadir@gmail.com

renovell@lirmm.fr

Yervant.Zorian@synopsys.com

Yervant.Zorian@synopsys.com

blanton@ece.cmu.edu

bcourtois@hotmail.co.uk

bozena@pultronics.com

dfsjsf4@clust.uib.es

lombardi@ece.neu.edu

erik.jan.marinissen@imec.be

michael.nicolaidis@imag.fr

isylla@ti.com

mike.ricchetti@Synopsys.com

harris@ics.uci.edu

zorian@Synopsys.com

rkapur@synopsys.com

magdy.abadir@gmail.com

sule.ozev@asu.edu

scott.davidson@oracle.com

bcourtois@hotmail.co.uk

Standards Working Groups

IEEE 1149.1

IEEE 1149.4

IEEE 1149.6

IEEE P1149.7

IEEE 1450-1999

IEEE 1450.1

IEEE 1450.2-2002

IEEE P1450.3

IEEE P1450.4

IEEE P1450.6-1

IEEE 1450.6-2

Christopher J. CLARK Intellitech Corporation - USA

Bambang SUPARJO Mentor Graphics - USA

Bill EKLOW - USA

Robert OSHANA Texas Instruments – USA

Gregory MASTON Synopsys, Inc. - USA

Tony TAYLOR

Gregg WILDER Texas Instruments - USA

Tony TAYLOR

Doug SPRAGUE IBM - USA

Jim O'REILLY Analog Devices - USA

Bruce CORY NVIDIA – USA

Saman ADHAM TSMC. – Canada

cclark@intellitech.com

bambang_suparjo@mentor.com

beklow56@gmail.com

roshana@ti.com

gmaston@synopsys.com

t.taylor@ieee.org

gwilder@ti.com

t.taylor@ieee.org

dsprague@us.ibm.com

jim_oreilly@ieee.org

bcory@nvidia.com

saman.adham@gmail.com

IEEE 1450.6-2005
IEEE P1450.7
IEEE 1500
IEEE 1532
IEEE 1581
IEEE 1687

IEEE P1838

Rohit KAPUR Synopsys, Inc. - USA
Jean-Louis CARBONERO STMicroelectronics - France
Yervant ZORIAN Synopsys. - USA
Neil JACOBSON Xilinx Corp. - USA
Heiko EHRENBERG GOEPEL Electronics - USA
Kenneth POSSE AMD - USA
Alfred CROUCH Asset InterTech - USA
Erik Jan MARINISSEN IMEC - Belgium

rkapur@synopsys.com
jean-louis.carbonero@st.com
zorian@synopsys.com
neil.jacobson@xilinx.com
h.ehrenberg@goepel.com
kepos@comcast.net
al.crouch@asset-intertech.com
erik.jan.marinissen@imec.be

IEEE EAST-WEST DESIGN & TEST SYMPOSIUM 2019 COMMITTEES

General Chairs

V. Hahanov
Y. Zorian – USA

General Vice-Chairs

R. Ubar – Estonia
P. Prinetto – Italy

Program Chair

S. Shoukourian –
Armenia
A. Ivanov – Canada

Program Vice-Chairs

Z. Navabi – Iran
M. Renovell – France

Finance Chairs

E. Litvinova

Publicity Chairs

S. Mosin – Russia
G. Markosyan –
Armenia

Public Relation Chair

V. Djigan – Russia

Steering Committee

V. Hahanov
R. Ubar – Estonia
Y. Zorian – USA

Organizing Committee

Z. Davitadze – Georgia
S. Chumachenko
E. Litvinova
A. Mishchenko

Program Committee

J. Abraham – USA
V. H. Abdullayev -
Azerbaijan
M. Adamski – Poland
A. S. Mohamed –
Egypt
A. Barkalov - Poland
R. Bazylevych
A. Chaterjee - USA
D. Devadze - Georgia
V. Djigan – Russia
A. Drozd
D. Efanov - Russia
E. Evdokimov
E. Gramatova -
Slovakia
G. Harutyunyan -
Armenia
A. Ivannikov – Russia
I. Kabin - Germany
M. Karavay - Russia
V. Kharchenko
M. Khalvashi - Georgia

K. Kuchukjan -
Armenia
V. Kureichik - Russia
W. Kuzmicz - Poland
A. Matrosova - Russia
V. Melikyan - Armenia
S. Mosin - Russia
O. Novak - Czech
Republic
A. Orailoglu - USA
Z. Peng - Sweden
A. Petrenko
N. Prokopenko -
Russia
J. Raik - Estonia
A. Romankevich
R. Seinauskas -
Lithuania
S. Sharshunov -
Russia
A. Singh - USA
J. Skobtsov
Z. Stamenkovic –
Germany
V. Tverdokhlebov -
Russia
V. Vardanian - Armenia
V. Yarmolik - Belarus

17th IEEE EAST-WEST DESIGN & TEST SYMPOSIUM (EWDTS 2019)

Batumi, Georgia, September 13-16, 2019

The main target of the IEEE East-West Design & Test Symposium (EWDTS) is to exchange experiences between scientists and technologies from Eastern and Western Europe, as well as North America and other parts of the world, in the field of design, design automation and test of electronic circuits and systems. The symposium is typically held in countries around East Europe, the Black Sea, the Balkans and Central Asia region. We cordially invite you to participate and submit your contributions to EWDTS'18 which covers (but is not limited to) the following topics.

- Analog, Mixed-Signal and RF Test
- ATPG and High-Level TPG
- Automotive Reliability & Test
- Built-In Self Test
- Debug and Diagnosis
- Defect/Fault Tolerance and Reliability
- Design Verification and Validation
- EDA Tools for Design and Test
- Embedded Software
- Failure Analysis & Fault Modeling
- Functional Safety
- High-level Synthesis
- High-Performance Networks and Systems on a Chip
- Internet of Things Design & Test
- Low-power Design
- Memory and Processor Test
- Modeling & Fault Simulation
- Network-on-Chip Design & Test
- Flexible and Printed Electronics
- Applied Electronics
Automotive/Mechatronics
- Algorithms
- Object-Oriented System Specification and Design
- On-Line Testing
- Power Issues in Design & Test
- Real Time Embedded Systems
- Reliability of Digital Systems
- Scan-Based Techniques
- Self-Repair and Reconfigurable Architectures
- Signal and Information Processing in Radio and Communication Engineering
- System Level Modeling, Simulation & Test Generation
- System-in-Package and 3D Design & Test
- Using UML for Embedded System Specification
- Optical signals in communication and Information Processing
- CAD and EDA Tools, Methods and Algorithms
- Hardware Security and Design for Security
- Logic, Schematic and System Synthesis
- Place and Route
- Thermal and Electrostatic Analysis of SoCs
- Wireless and RFID Systems Synthesis
- Sensors and Transducers
- Medical Electronics
- Design of Integrated Passive Components

The Symposium will take place in Batumi – is the second-largest city of Georgia, located on the coast of the Black Sea in the country's southwest. It is situated in a subtropical zone near the foot of the Lesser Caucasus Mountains. Much of Batumi's economy revolves around tourism and gambling, but the city is also an important sea port and includes industries like shipbuilding, food processing and light manufacturing. Since 2010, Batumi has been transformed by the construction of modern high-rise buildings, as well as the restoration of classical 19th-century edifices lining its historic Old Town.

CONTENTS

CFI: Control Flow Integrity or Control Flow Interruption? Nicolo Maunero, Paolo Prinetto, Gianluca Roascio	1
From Abstract Modeling of ADAS Applications to an Accelerator-based Hardware Realization Samira Ahmadi Farsani, Katayoon Basharkhah, Amin Mohaghegh, Zainalabedin Navabi	7
Unified STIL Flow: A Test Pattern Validation Approach for Compressed Scan Designs Slimane Boutobza, Andrea Costa, Sorin Popa	13
An Accelerator-based Architecture Utilizing an Efficient Memory Link for Modern Computational Requirements Saba Yousefzadeh, Katayoon Basharkhah, Nooshin Nosrati, Rezgar Sadeghi, Jaan Raik, Maksim Jenihhin, Zainalabedin Navabi	23
Antenna Array Calibration Algorithm Based on Phase Perturbation Victor Djigan, Vladislav Kurganov	29
Power Supply Noise Rejection Improvement Method in Modern VLSI Design Vazgen Melikyan, Artur Mkhitarian, Hakob Kostanyan, Hayk Grigoryan, Harutyun Kostanyan, Mushegh Grigoryan, Ruben Musayelyan, Hayk Margaryan	34
Making System Level Test Possible by a Mixed-mode, Multi-level, Integrated Modeling Environment Nooshin Nosrati, Katayoon Basharkhah, Rezgar Sadeghi, Carna Zivkovic, Christoph Grimm, Zainalabedin Navabi	38
Fast and Efficient Implementation of Lightweight Crypto Algorithm PRESENT on FPGA through Processor Instruction Set Extension Abdullah Varici, Gurol Saglam, Seckin Ipek, Abdullah Yildiz, Sezer Gören, Aydin Aysu, Deniz Iskender, T. Baris Aktemur, H. Fatih Ugurdag	43
Qubit Test Synthesis Processor for SoC Logic Wajeb Gharibi, David Devadze, Vladimir Hahanov, Eugenia Litvinova, Ivan Hahanov	48
Increasing the Effective Volume of Digital Watermark Used in Monitoring the Program Code Integrity of FPGA-Based Systems Kostiantyn Zashcholkin, Oleksandr Drozd, Ruslan Shaporin, Olena Ivanova, Yulian Sulima	53
Unit Regression Test Selection According To Different Hashing Algorithms Hakobyan Hovhannes H., Vardumyan Arman V., Kostanyan Harutyun T.	59
SCOAP-based Directed Random Test Generation for Combinational Circuits Seyyede Maryam Ghasemy, Maryam Rajabalipanah, Saeideh Sarmadi, Zainalabedin Navabi	63

OR2-NOC: Offline Robust Routing Algorithm For 2-D Mesh Nocs Architectures Arezoo Beheshti Soofian, Mina Zolfy Lighvan, Zahra Eghbali	68
Simulation of Nodes and Blocks of Matching Processor of the Parallel Dataflow Computing System "Buran" Nikolay Levchenko, Anatoly Okunev, Dmitry Zmejev	73
Optimized Time-Delayed Feedback Control of Fractional Chaotic Oscillator with Application to Secure Communications Amir Rikhtegar Ghiasi, Mona Saber Gharamaleki, Elaheh Mohammadi asl Khasraghi, Zahra Sattarzadeh Kalajahi	78
Implementation Variants of the Global Distributed Associative Computing Environment for the Parallel Dataflow Computing System "Buran" Nikolay Levchenko, Anatoly Okunev, Dmitry Zmejev	84
Researching Resilience a Holistic Approach Zoya Dyka, Ievgen Kabin and Peter Langendörfer	88
Modeling and Debugging Tools Development for Recurrent Architecture Dmitry Khilko, Yury Stepchenkov, Yury Shikunov, George Orlov	92
Caution: GALS-ification as a Means against SCA Attacks Zoya Dyka, Ievgen Kabin, Dan Klann, Frank Vater and Peter Langendoerfer	97
Unit Regression Test Selection Mechanism Based on Hashing Algorithm Melikyan Vazgen Sh., Hakobyan Hovhannes H., Kaplanyan Taron K., Momjyan Arsen M.	103
Qubit Fault Detection in SoC Logic Mikhail Karavay, Vladimir Hahanov, Eugenia Litvinova, Hanna Khakhanova, Irina Hahanova	108
The Fault Tolerant CMOS Logical C-Element for Digital Devices Resistant to Single Nuclear Particles Yuri V. Katunin, Vladimir Ya. Stenin	115
Development of a Simulation Tool to Estimate Final Electricity Consumption and Determine the Optimum Cooling System for Data Centers Beyzanur Toprak, Beyzanur Bora and Gül Nihal Güğül	119
Terms of Arrangement Reckoning Self-Checking Embedded Check Circuits Based on Boolean Complement up to Constant-Weight Code '1-out-of-3' Dmitry Efanov, Valery Sapozhnikov, Vladimir Sapozhnikov, German Osadchy, Dmitry Pivovarov	125
Use of Natural Information Redundancy in On-Line Testing of Computer Systems and their Components Oleksandr Drozd, Anatoliy Sachenko, Svetlana Antoshchuk, Julia Drozd, Mykola Kuznietsov	131

Self-Dual Complement Method up to Constant-Weight Codes for Arrangement of Combinational Logical Circuits Concurrent Error-Detection Systems Dmitry Efanov, Valery Sapozhnikov, Vladimir Sapozhnikov, German Osadchy, Dmitry Pivovarov	136
Decision Making in VLSI Components Placement Problem Based on Grey Wolf Optimization Elmar V. Kuliev, Vladimir VI. Kureichik, Ilona O. Kursitys	144
Intelligent Flow Meter on Acoustic Multivibrator Zh. A. Sukhinets, Gulin A. I., Bureneva O.I., Prokopenko N.N., Valiamova O.O.	148
Technique to Simulate Oscillator Circuits with the Degradation Models Mark M. Gourary, Sergey G. Rusakov, Sergey L. Ulyanov, Michael M. Zharov	153
Polynomial Code with Detecting the Symmetric and Asymmetric Errors in the Data Vectors Ruslan B. Abdullaev, Dmitrii V. Efanov, Valerii V. Sapozhnikov, Vladimir V. Sapozhnikov	157
Diagnostics of Audio-Frequency Track Circuits in Continuous Monitoring Systems for Remote Control Devices: Some Aspects Dmitrii V. Efanov, German V. Osadchy, Valerii V. Khóroshev, Dmitrii A. Shestovitskiy	162
Length Limiting of Quantum Key Distribution at Two-Stage Synchronization Rumyantsev K.E., Shakir H.H.Sh.	171
Technological Foundations of Traffic Controller Data Support Automation Joseph M. Kokurin, Dmitrii V. Efanov	176
Sum Codes of Weighted Data Bits for Objectives of Automation Logical Devices Technical Diagnostics Dmitry Efanov, Valery Sapozhnikov, Vladimir Sapozhnikov, German Osadchy, Teng Teng	181
Algorithm for Extraction of the Iris Region in an Eye Image Sh.Kh. Fazilov, O.R. Yusupov	191
Processing An Effective Method For Clock Tree Synthesis Narek Avdalyan, Kamo Petrosyan	196
Development of Automation Systems at Transport Objects of MegaCity Andrei Belyi, Dmitrii Shestovitskii, Valerii Myachin, Dmitrii Sedykh	201
Advanced Indication of the Self-Timed Circuits Yury Stepchenkov, Yury Shikunov, Yury Diachenko, Denis Diachenko, Yury Rogdestvenski	207

Main Solutions of Structural Health Monitoring in Managing the Technical Condition of Transport Objects Andrei Belyi, Dmitrii Shestovitskii, Eduard Karapetov, Dmitrii Sedykh, Vladimir Linkov	213
A Technique for Semiconductor Devices Modeling Using Physical Templates Alexandr M. Pilipenko, Vadim N. Biryukov, Alexander I. Serebryakov	219
Statistical Analysis of Discriminators under the Influence of Additive Correlated non-Gaussian Noise Described by Markov Processes Artyushenko V.M., Volovach V.I.	223
Accurate Soft Error Rate Reduction using Modified Resolution Method Alexander Stempkovskiy, Dmitry Telpukhov, Vladislav Nadolenko	229
Neural Net as Pseudo-Inverse Filter in Speech Coding Problem Rustam Latypov, Evgeni Stolov	235
Protograph Sieving Method for Construction Moderate Length Multi-Edge Type QC-LDPC Codes Svistunov German, Usatyuk Vasiliy, Egorov Sergey	239
Calculating the Parameters of the Short-Range Microwave Information Channel Resistant to Signal Fading Artyushenko V. M., Volovach V. I.	243
The Implementation of the Genetic Algorithm Using Cloud-Based Computing on the Internet Kureichik V. M., Logunova J.A.	248
A Template Model of Junction Field-Effect Transistors f or a Wide Temperature Range Alexandr M. Pilipenko, Vadim N. Biryukov, Nikolay N. Prokopenko	252
Synthesis of Signal Quadrature Processing Algorithms under the Influence of Band-limited non-Gaussian Noise Artyushenko V.M., Volovach V.I.	256
Planar Butler Matrix Based on Compact Taps Denis A. Letavin	260
All-Pass Second-Order Active RC-Filter with Pole Q-Factor's Independent Adjustment on Differential Difference Amplifiers Darya Yu. Denisenko, Nikolay N. Prokopenko, Nikolay V. Butyrlagin	263
Planar Compact Directional Coupler on Artificial Transmission Lines Denis A. Letavin	267
Silicon Photomultipliers' Analog Interface with Wide Dynamic Range Oleg V. Dvornikov, Yaroslav D. Galkin, Nikolay N. Prokopenko, Alexey E. Titov, Vladimir A. Tchekhovski, Anna V. Bugakova	270

A Novel Technique For Atomic Instructions Functional Verification Using Lock Contention Analysis Chibisov Peter, Grevtsev Nikita	274
Calculation of Phase Center of Arbitrary Electromagnetic Radiation Sources in Near Field Zone Nikolay Anyutin, Ivan Malay, Alexey Malyshev	281
Method of Calculating the Spare Parts System Availability for Electronic Means Pankovsky B. E., Polesskiy S. N.	285
Ternary Questionnaires Dmitrii V. Efanov, Valerii V. Khóroshev	289
Harmonic Distortions in Analog Interfaces Based on Differential Difference Amplifiers Nikolay V. Butyrlagin, Anna V. Bugakova, Nikolay N. Prokopenko, Mikhail F. Mitsik, Alexey E. Titov	301
Intelligent Sensor Measurement of GTE Gas Temperature with Thermistors Zh. A. Sukhinets, A. I. Gulin, N. M. Safyannikov, N. N. Prokopenko, O. O. Valiamova, O.I. Bureneva	305
Application of Modern Microelectronic Technology in Marshalling Process of Railway Stations Michael A. Gordon, Alexey N. Kovkin, Dmitry V. Sedykh, Anton A. Movshin, Oleg A. Abramov	309
Ternary Parity Codes: Features Dmitrii V. Efanov	315
Fast and Secure Unified Field Multiplier for ECC Based on the 4-Segment Karatsuba Multiplication Ievgen Kabin, Zoya Dyka, Dan Klann and Peter Langendoerfer	320
Construction of Length and Rate Adaptive MET QC-LDPC Codes by Cyclic Group Decomposition Usatyuk Vasiliy, Egorov Sergey, German Svistunov	326
A Technique for the Accounting of Surrounding Circuitry During Generation of the Simplified Models Mark M. Gourary, Sergey G. Rusakov, Sergey L. Ulyanov, Michael M. Zharov	331
A Signal Processing Approach for the Failure Analysis of Rolling-Element Bearing of Vehicle Brake Tester Used at a Vehicle Inspection Station Selman Kulac	337
Modified Modular Unit Bits Sum Codes with Arbitrary Account Modules Dmitrii V. Efanov, Anna O. Filippochkina, Mariia V. Ivanova	343

Systems for Reflectometry Analysis of Defects in Metal Structures of Transport Mobile Objects Julia V. Alevetdinova	350
Parametric Optimization Subsystem in LTspice Environment of Analog Microcircuits for Operation at Low Temperatures Maxim V. Liashov, Nikolay N. Prokopenko, Andrei A. Ignashin, Oleg V. Dvornikov, Alexey A. Zhuk	356
Boosting Model of Bioinspired Algorithms for Solving the Classification and Clustering Problems Ilona Kursitya, Alexander Natskevich, Elvira Tsyruunikova	360
Modeling Technique of Large Signal Dynamics for Electromagnetic Levitation Melting System Idan Sassonker, Moria Elkayam, Alon Kuperman	366
On a Method for Segmentation of Memory Instances with Row Redundancies Karen Amirkhanyan, Valery Vardanian	371
Elaboration of the Functioning Algorithm of Three-Dimensional Model of Computer System Safety Victor V. Zhilin, Larissa V. Cherckesova, Irina I. Drozdova, Vitaliy M. Porksheyan, Ivan A. Sakharov, Olga A. Safaryan, Andrey G. Lobodenko, Sergey A. Morozov	376
Interface and Software for the System of Automatic Seeding of Grain Crops Maksim A. Litvinov, Maksim N. Moskovskiy, Ilya V. Pakhomov, Igor G. Smirnov	380
Cross-Platforming Web-Application of Electronic On-line Voting System on the Elections of Any Level Evgeniy V. Palekha, Olga A. Safaryan, Irina S. Trubchik, Vitaliy M. Porksheyan, Olga N. Manaenkova, Sergey A. Morozov, Larissa V. Cherckesova, Boris A. Akishin	384
Theoretical Bases of the Course Motion Two Axles Agriculture Transports Vehicle According Wheels Slipping Maksim A. Litvinov, Maksim N. Moskovskiy, Ilya V. Pakhomov, Anatoly A. Gulyaev	388
Modification and Optimization of Solovey-Strassen's Fast Exponentiation Probabilistic Test Binary Algorithm Nikita Ye. Myzdrikov, Olga A. Safaryan, Ivan Ye. Semeonov, Irina V. Reshetnikova, Vasiliy I. Yukhnov, Andrey G. Lobodenko, Larissa V. Cherckesova, Vitaliy M. Porksheyan	392
Reliability Issues in the Parallel Dataflow Computing System Nikolay Levchenko, Anatoly Okunev, Dmitry Zmejev	395
Modification and Optimization of Pollards's Factorization p-Method by Means of Recursive Algorithm of Number Calculation Factorization Ivan A. Smirnov, Larissa V. Cherckesova, Pavel V. Razumov, Yelena A. Revyakina, Nickolay V. Boldyrikhin, Vitaliy M. Porksheyan, Olga A. Safaryan, Andrey G. Lobodenko	400

Deriving Low Power Test Sequences Detecting Robust Testable PDFs A. Matrosova, V. Andreeva, V. Tychinskiy	406
Development of Modified Block Cipher Algorithm TEA, Free from Vulnerability of “Connected Keys Attack” Sergey A. Klyokta, Alexander I. Zhukov, Nikita I. Chesnokov, Larissa V. Cherckesova, Irina A. Pilipenko, Olga A. Safaryan, Andrey G. Lobodenko, Vitaliy M. Porksheyana	410
Masking Internal Node Faults and Trojan Circuits in Logical Circuits A. Matrosova, V. Provkina, E. Nikolaeva	416
Masking Robust Testable PDFs Anzhela Matrosova, Sergei Ostanin, Semen Chernyshov	420
Associative Processors: Application, Operation, Implementation Problems Egor Kuzmin, Nikolay Levchenko, Anatoly Okunev	424
SWIELD: An In Situ Approach for Adaptive Low Power and Error-Resilient Operation Mitko Veleski, Rolf Kraemer, Milos Krstic	430
Automating of Human Resources Management using Genetic Algorithms Agata V. Markevich, Valentina G. Sidorenko	436
Evaluating the Length of Distinguishing Sequences for Non-Deterministic Input/Output Automata Igor Burdonov, Alexandr Kossachev, Nina Yevtushenko, Alexey Demakov	445
Voltage Regulation Analysis in Energy Transmission Systems Using STATCOM Hamza Feza Carlak, Ergin Kayar	450
Research of the Effect of Discrete Light Sources on Seeds of Vegetable and Green Cultures and the Possibility of their Approximation to Modified Sunlight Danila Yu. Donskoy, Alexander D. Lukyanov, Marko Petković, Eugenia P. Kluchka	457
Deriving Adaptive Homing Sequences for Weakly Initialized Nondeterministic FSMs Evgenii Vinarskii, Aleksandr Tvardovskii, Larisa Evtushenko, Nina Yevtushenko	461
Non-Canonical Topography of the z-Plane Discretized due to Quantization of the IIR Digital Filter Coefficients Vladislav Lesnikov, Tatiana Naumovich, Alexander Chastikov	466
Modification of the U-Net Neural Network in the Task of Multichannel Satellite Images Segmentation Vladimir Khryashchev, Roman Larionov, Anna Ostrovskaya, Alexander Semenov	470
Complementary JFETs Integrated into the Microwave Complementary Bipolar Double Self-Aligned Technology Dmitry G. Drozdov, Nikolay N. Prokopenko, Evgeny M. Savchenko, Andrey I. Grushin, Pavel A. Dukanov	474

The Discrete Structure of the Zeros and Poles Location in the z-Plane of the Arbitrary Order IIR Digital Filters with a Finite Word Length Vladislav Lesnikov, Tatiana Naumovich, Alexander Chastikov, Alexander Metelyov	478
Permanent Monitoring Systems of the Contact-Wire of Railroad Catenary: the Main Tasks of Implementation Dmitrii V. Efanov, German V. Osadchy, Dmitrii V. Barch, Andrei A. Belyi	484
Design of Real-Time System Logic Control on FPGA Maryna Miroshnyk, Dariia Rakhlis, Inna Filippenko, Elvira Kulak, Maksym Hoha, Mykyta Malakhov, Vladyslav Sergienko	488
Emerging Culture of Social Computing Anastasia Hahanova, Svetlana Chumachenko, Vladimir Hahanov, Abdullayev Vugar Hacimahmud, Ka Lok Man, Alexander Mishchenko	492
Forest Areas Segmentation on Aerial Images by Deep Learning Vladimir Khryashchev, Anna Ostrovskaya, Vladimir Pavlov, Roman Larionov	497
An Analysis of LockerGoga Ransomware Alexander Adamov, Anders Carlsson, Tomasz Surmacz	502
An Analysis of Sampling Effect on the Absolute Stability of Discrete-time Bilateral Teleoperation Systems, Amir Aminzadeh Ghavifekr, Seyedshahab Chehraghi, Giacomo De Rossi	507
The Software Platform for Evaluation of Effectiveness of Network Systems Analysis Technologies Olha Ponomarenko, Valeriy Gorbachov, Abdulrahman Kataeba Batiaa, Oksana Kotkova	513
Multidimensional Hierarchical Model of Behavioral Check of Distributed Information Systems Oleksandr Martynyuk, Oleksandr Drozd, Hanna Stepova, Dmitry Martynyuk and Lyudmila Sugak	517
Development of Method For Automation of SPICE Models Generation Melikyan Vazgen Sh., Martirosyan Meruzhan K.	523
Comparison of Grapheme-to-Phoneme Conversions For Spoken Document Retrieval Dmitriy Prozorov, Alexandra Tatarinova	527
Formalized Methods of Analysis and Synthesis of Electronic Document Management of Technical Documentation Dilshod Baratov, Aripov Nazirjon and Ruziev Davron	531
Non-Invasive System for Determining the Level of Iron in the Blood Andrey Azarov, Elena Shirokova, Igor Shirokov	540

The Using of Electronic Document Management Tools of Technical Documentation for the Assessment of the Life of the Train Traffic Control Devices Dmitry V. Sedykh, Michael N. Vasilenko, Andrei Belyi, Denis V. Zuyev, Michael A. Gordon	544
Automation of Layout Design of Spiral Conical Scans Marina Byrdina, Lema Bekmurzaev, Mikhail Mitsik, Dmitry Kelekhsaev and Anatoly Kondratenko	548
Secure Communication Using the Synchronization of Time-Varying Complex Networks by Fuzzy Impulsive Method Reza Behinfaraz, Sehraneh Ghaemi, Sohrab Khanmohammadi, Mohammadali Badamchizade	552
Description of the Spatial Shape Surface of an Air Supported Dynamic Figure Mikhail Mitsik, Lema Bekmurzaev, Marina Byrdina, Olga Aleynikova, Victor Kokhanenko	556
Solution of the Dynamic Problem of Optimal Design of Electronic Devices Based On the Gravity Center Method Mikheil Donadze and Zurab Meskhidze	560
Intelligent Transport Systems as a Way to Improve the Quality of the Rail-Truck Multimodal Freight Transportation Natalia Goncharova	565
Remote Administration of Information Systems Via E-mail Zaza Davitadze, Gregory Kakhiani and George Beria	572
Method of Indirect Steganographic Embedding Based on Functionality for Adaptive Position Number Vladimir Barannik, Dmitry Barannik, Nataliy Barannik	577
Surface visualization of flexible elastic shells Marina V. Byrdina, Lema A. Bekmurzaev, Mikhail F. Mitsik, Svetlana V. Rubtsova	582
AUTHORS INDEX	586

CFI: Control Flow Integrity or Control Flow Interruption?

Nicolò Maunero
CINI Cybersecurity National Lab.
Turin, Italy
nicolo.maunero@consorzio-cini.it

Paolo Prinetto
DAUIN - Politecnico di Torino
CINI Cybersecurity National Lab.
Turin, Italy
paolo.prinetto@polito.it

Gianluca Roascio
CINI Cybersecurity National Lab.
Turin, Italy
gianluca.roascio@consorzio-cini.it

Abstract—Runtime memory vulnerabilities, especially present in widely used languages as C and C++, are exploited by attackers to corrupt code pointers and hijack the execution flow of a program running on a target system to force it to behave abnormally. This is the principle of modern Code Reuse Attacks (CRAs) and of famous attack paradigms as Return-Oriented Programming (ROP) and Jump-Oriented Programming (JOP), which have defeated the previous defenses against malicious code injection such as Data Execution Prevention (DEP). Control-Flow Integrity (CFI) is a promising approach to protect against such runtime attacks. Recently, many CFI solutions have been proposed, with both hardware and software implementations. But how can a defense based on complying with a graph calculated *a priori* efficiently deal with something unpredictable as exceptions and interrupt requests? The present paper focuses on this dichotomy by analysing some of the CFI-based defenses and showing how the unexpected trigger of an interrupt and the sudden execution of an Interrupt Service Routine (ISR) can circumvent them.

I. INTRODUCTION

Computing devices are nowadays a corner stone of our daily life. Almost all services have now been translated into digital, and even the objects that surround us are computer-controlled and mutually connected, in what is usually referred to as the Internet of Everything. In such a scenario, ensuring security of data and privacy has become increasingly important.

Securing this new global network concerns not only the integrity and the trustworthiness of links and interconnections, but also relates to aspects strictly bound to embedded systems, such as the adopted programming languages. Most lines of code are still written in C and C++ [2], since these languages permit a good degree of low-level control without losing the advantages of high-level statements. Although, the possibility of operating at low level can turn into a disadvantage when dealing with security issues, since the direct management of memory pointers opens the door to a wide range of vulnerabilities. These include, among others, *dangling pointers* [4], i.e., pointers to live objects which are mistakenly freed and can be corrupted during the execution, and *buffer overflows* [36], i.e., out-of-bounds writes of a memory buffer which corrupts adjacent data on stack or heap.

Some of these vulnerabilities may also enable corruption of *code pointers*, used as argument of indirect control-flow transfer instructions. Malicious attackers, by tampering with

them, succeed in taking full control over the program execution path. In such attacks, rather than by injecting code, a malware is executed by redirecting the flow of the program to portions of code that already exist in memory but are not meant to be executed in that order. This is the fundamental aspect of *code-reuse attacks* and famous exploit paradigms such as *Return-Oriented Programming* (ROP) [41] [12] [14] [37] and *Jump-Oriented Programming* (JOP) [9] [17]. Attackers individuate, within the code, short sequences of instructions (typically from 2 to 5) called *gadgets*. A gadget always ends with an indirect control-flow transfer instruction which can be used as a trampoline for the next gadget. By opportunely selecting gadgets and chaining them together, it is possible to force very dangerous behaviours, with Turing-complete compute capabilities [41] [46].

In [3], the enforcement of the *Control-Flow Integrity* (CFI) as basic defense was formalized. CFI dictates that, during program execution, whenever a control-flow transfer occurs, it must target a valid destination, as determined by a *Control-Flow Graph* (CFG) created at compile time. Several solutions for the CFI have been proposed [8] [22] [48] [35] [50] [30] [23] [42] [19] [21] [49]. Commonly, two phases are distinguished: during the *offline* phase, the intended flow transfers of the program are computed, and in the *online* phase it is verified whether these transfers are respected by the running program without divergencies. The offline phase is usually performed resorting to a static analysis of the program binary, finding its *basic blocks*. As in [32], a basic block is defined as a linear sequence of program instructions having one entry (the first instruction executed) and no branches out except at the exit (a control-flow transfer instruction). All statements within a basic block are executed before transferring the control to the next basic block. Transfers and basic blocks assume the identity of edges and vertices within the Control-Flow Graph, and if an online monitor (software or hardware) is able to guarantee that the program does not take paths different from those established in such a graph, then the program is considered secure and immune to redirection attacks.

The CFG represents the intended behaviour of the program, but actually can not deal with unpredictable events that may happen at runtime. In real cases, interrupt requests can be sent to the processor at any time. Serving a request is an exceptional

flow transfer, not triggered by any instruction, which preempts the execution even in the middle of a basic block, and forces the control to move to the Interrupt Service Routine (ISR) location. Such routines: (a) save the context of the current execution (i.e., the instruction pointer and the status word, as well as the used registers) on the stack, (b) acknowledge the request, (c) at their completion, they restore the context and return the control to the application. It worth pointing out that they contain normal code as any other function, including, for instance, possible local buffers that can corrupt the stack if overflowed or any other memory vulnerability. ISRs could thus be used to start a control-flow-hijacking attack, with the significant difference that, in these cases, all the static defense techniques that rely on the CFG enforcement, fail. During the offline analysis phase, both the locations from which ISRs are activated and, consequently, the return locations, are unknown. As a consequence, there is no way to monitor and protect them resorting to CFGs.

When the application runs on top of an Operating System (OS), the response to an interrupt request is demanded to the kernel. Programmers that intend to adopt a CFI enforcement solution for their programs are forced to rely on the OS capabilities to prevent possible problems related to interruptions. When instead no OS is present (bare-metal), the Interrupt Vector Table (IVT) and the ISRs are totally part of the program, so the above mentioned issues must be carefully addressed.

The present paper aims at showing how some of the classic examples of CFI enforcement, either hardware-assisted or purely software-implemented, fail in their protection purposes under the presence of hardware interrupts and vulnerable ISRs. The rest of the paper is organised as follows: Section II provides some technical background on code-reuse attacks and common solutions, while Section III analyses in details some CFI solutions and explains why and how they are vulnerable in presence of interrupts. Section IV concludes the paper.

II. BACKGROUND

In computer security, the term *Arbitrary Code Execution* (ACE) is commonly used to describe the ability to execute arbitrary commands or code on an attacked machine. ACE is achieved through tampering with the *instruction pointer* (in some architecture referred to as *Program Counter*) of a running program. The instruction pointer points to the next instruction to be executed, therefore by controlling its value an attacker can control the instruction to be executed next. To execute arbitrary code, attackers exploit possible memory vulnerabilities [36] [4] [44] present in a program to redirect the instruction pointer to malicious code, often referred to as *payload*.

Traditionally, the payload was injected together with the corrupted instruction pointer in the memory of the program (*Code Injection*) thank to vulnerabilities typically present in the stack [33]. Such exploits were made impossible after the wide adoption of *Data Execution Prevention* (DEP) [43] and *Write XOR Execute* policy [45], for which no memory location

can be both writable (W) and executable (X). Attackers then reacted by devising a new attack paradigm, in which the payload is composed of code already present in the memory image of the application under attack. This was the born of the so called *Code Reuse Attacks* (CRA). The standard C library, *libc*, is the usual target, since it is loaded in nearly every program. By carefully arranging values on the stack, an attacker can cause a sequence of functions to be invoked, one after the other, with arbitrary arguments (*Return-into-libc*, [1]). DEP-based defences are thus circumvented, but still with some limitation, since fully arbitrary execution cannot be reached.

In [41], the authors stated that “*in any sufficiently large body of executable code there will exist sufficiently many useful code sequences that an attacker who controls the stack will be able [...] to cause the exploited program to undertake arbitrary computation*”. This is the idea behind the exploit known as *Return-Oriented Programming* (ROP). ROP is based on the assumption that return addresses on the stack can point anywhere, not just to the beginning of functions. Therefore, the control flow can be hijacked through a series of small sequences of instructions, each ending with a **ret**, known as *gadgets*. In a large enough codebase (such as *libc*), there is a massive selection of gadgets to choose from, and the attackers achieve the maximum of expressiveness [46]. On the x86 platform, the attack is made stronger by the fact that, since there is no fixed instruction length, any sequence of raw bytes can be interpreted as an instruction, and the rogue return address can point even in the middle of an opcode transforming it into another.

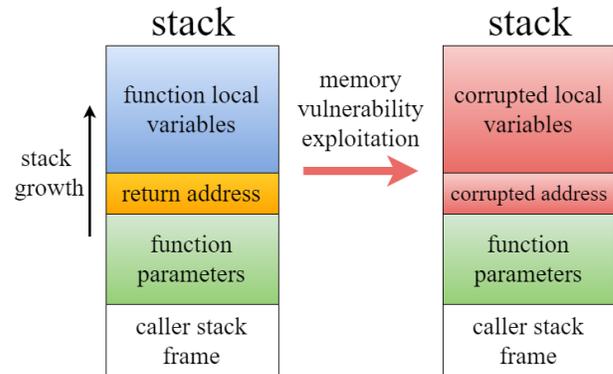


Fig. 1. Return address corruption, start point of Return-into-libc and ROP attacks.

The concept of ROP was first generalised to other architectures [12] [25] [15] [14] [31] and then extended to non-**ret**-ended gadgets: **ret** is useful in gadgets as it transfers the control flow using a program value (the return address on top of stack), not precalculated at compile time. As a result, indirect formats of **jmp** and **call** can as well be used to reach a desired instruction sequence. The concepts of *Jump-Oriented Programming* (JOP) [9] [17], *Call-Oriented Programming* (COP) [39], and others [40] [29] were introduced.

In the last years, research community and companies started elaborating and adopting different types of solution to counter

CRA. *Address Space Layout Randomisation* (ASLR) [7] is a countermeasure taken at link-time which randomises the memory layout of the application, making it harder for an attacker to know the exact address of libraries code. Actually, in 32-bit architecture the introduced entropy is too low, and brute-force attacks can easily break the defense. Furthermore, it suffers of information disclosure, since just the base address of each segment is randomised, and therefore gaining the knowledge of a single address leads to compute the library segment base address in a straightforward manner [38].

In [20], the concept of *stack canary* or *stack cookie* was introduced: when a function is called, an additional word with a known value can be pushed on top of the stack, which is placed between the return address and the local variables. When the function returns, the value of the canary is checked, and, if it is found changed, the program is considered under attack and terminated. The canary can have a random value difficult to guess or can be composed of terminator characters, making it difficult to manipulate using input function (such as `gets()`), since terminator character breaks the input streams when recognised. However, canaries have been shown to be circumventable with more targeted stack-smashing attacks [5].

In order to address the stack smashing problem as-a-whole, a *Shadow Call Stack* (SCS) can be used [47] [26] [19] [11] [10]. Basically, at call-time, the return address is both saved on top of the normal stack and on top of an additional shadow one, accessible only by the processor in a private manner. At return time, the instruction pointer is **poped** from both stacks, and the values compared. If a mismatch is found, an exception is raised. Even if this solution protects the stack, it is not sufficient to fully protect an application, as it only blocks stack smashing and does not address memory vulnerability present in other segments (heap, bss, data, etc), with the consequence that exploits such as JOP can be easily performed.

Heuristic-based approaches claim to detect CRAs by typically monitoring the number of branches of the program and block it when suspicious behaviour is sensed. The assumption is that gadgets for ROP and JOP attacks usually consist of no more than 5 instructions. DROP [16], kBouncer [34] and ROPecker [18] are examples of heuristic-based defenders. However, it has been demonstrated that the heuristic can be easily thwarted by executing, between malicious jumps, longer sequences of non-jumping instructions or branches considered as secure [28].

Solutions presented so far can still be valid mitigation techniques, relatively simple to implement, but each of them addresses the problem of code redirection attacks just with respect to one of the vulnerabilities that lead to the exploit, without an all-encompassing vision. The paper [3] first tried to change perspective by introducing the concept of *Control-Flow Integrity* (CFI) as basic defense against CRAs, regardless of the vulnerability that may cause them. The concept behind CFI is monitoring the program at runtime to detect abnormal diversion from what is stated in its *Control-Flow Graph*. Each node in the CFG represents a *basic block*, which is a group of non-jumping instructions executed sequentially.

Edges represent *branches* in the control flow, caused by jump, call, or return instructions. The CFG is defined before the execution, through a static analysis of the source code or of the binary, or by *execution profiling*, a test run which creates the possible paths. Then, at runtime, the dynamic control flow changes are restricted to the static CFG. Typically, just *indirect* formats of branching instructions are monitored, as it is usually assumed that the code is immutable, not self-modifying and not generated just-in time.

CFI policies are clustered into *coarse-grained* if the monitoring is not done by strictly enforcing the CFG, but based on simple rules, such as ensuring that **ret** targets are preceded by a **call**, or indirect calls only target prologues of functions, and similar. *Fine-grained* policies, instead, check that the execution traverses valid edges of the pre-computed CFG, only. However, coarse-grained solutions are not so different from heuristics, as both aim at distinguishing the rogue behaviours from those that *most probably* are benevolent. But *most probably* does not mean *certainly*, especially when we are dealing with clever attackers. Recent works [24] [27] show how it is possible to induce such security policies to believe that actions are within the rules when they are not. In [13] the authors showed that just 70 KB of binary code retrieved in 10 different executables within `/usr/bin` of Linux have been sufficient for mounting fully call-preceded ROP attacks.

Therefore, fine-grained CFI solutions are the only CFI policies ensuring that all control flow transfers within a program are only the ones intended by the design. In the next Section, we will compare the fine-grained CFI strategy and its most well-known implementations with the problem of unplanned interrupts, and we will show how defences can be bypassed.

III. MAIN ISSUES OF COMMON SOLUTIONS

A. Control-Flow Graph

As defined in [6], the *Control-Flow Graph* (CFG) is a directed graph in which the nodes represent the basic blocks (see Section I) of a program and the edges represent the control-flow transfers. The CFG is a good instrument for outlining the behaviour of a program, but it is not sufficient to completely describe the runtime execution of a program, in particular considering *interrupts*, that can occur at any time and are thus absolutely unpredictable.

When an interrupt is triggered, a pair of “phantom” edges are created, outside the CFG: the former one connects the just-executed instruction to the initial basic block of the ISR, while the latter one connects the exit point of the ISR to the instruction following the interruption site. At offline analysis time, it is not possible to know where these “phantom” edges will be located. Moreover, their occurrence is not just untraceable, but against the definition of CFG, too, as they can start from (and arrive to) an instruction internal to a basic block.

The processor, before serving an interrupt request, saves the context of the currently executed program in order to be able to restore it when the execution is resumed. The problem is that the code of the ISR is a just another piece of

code, not immune to vulnerabilities from which an attack can start. In particular, if the routine contains one of the memory vulnerabilities presented above, the return address may be corrupted and an attacker may gain control over the program execution redirecting it to potentially dangerous code.

Given the intrinsic asynchronous nature of interrupts, no static analysis can provide a valid mean to monitor this type of transfers at runtime.

B. Binary Instrumentation

The presence of interrupts is an issue not just for its non-traceability, but also for the fact that it can break the defences based on fine-grained CFI. To achieve it, several binary instrumentation approaches have been proposed.

One of these is the *label-based binary instrumentation*, first introduced by Abadi *et al.* [3] in their milestone paper. The label-based approach relies on modifying the compiled binary to insert unique IDs at the beginning of each basic block. Before each indirect branch, few instructions are inserted to check if the destination basic block's ID is in fact targeted by the instruction. Control flow tampering causes the check to fail, since the destination label ID will not match the label ID stored inside the program. Anyway, attacks are still possible if an interrupt request is served in the middle of the code used to instrument the jump.

The following code

```
cmp [ecx], 12345678h
jne violation
lea ecx, [ecx+4]
jmp ecx
```

is used to ensure that the `jmp ecx` at the bottom reaches the code starting with that ID, such as

```
.data 12345678h
mov eax, [esp+4]
...
```

Anyway, let us suppose that, thank to a memory vulnerability, an attacker has already tampered with the content of `ecx` to exploit that indirect jump. If an interrupt request is served between the `cmp` and the `jne` that triggers the violation, the processor status word (PSW) is pushed with the return address on top of the stack. The ISR may contain a vulnerability, and the PSW may be maliciously overwritten, such that the ZF flips and, when returning, the violation is bypassed and the desired piece of code is reached by the attacker.

The following instructions

```
mov eax, 12345677h
inc eax
cmp [ecx+4], eax
jne violation
jmp ecx
```

represents an alternative way to instrument the code if the destination is instrumented as follows

```
prefetchnta [12345678h]
mov eax, [esp+4]
...
```

using a side-effect-free x86 prefetch instruction. This version is even more exposed to interrupt issues: the previous exploitation is still possible if a vulnerable ISR is executed between the `cmp` and the `jne`, but another attack is possible. The `cmp` with the ID resorts to a register instead of an immediate. Therefore, let us suppose that a vulnerable ISR is triggered after the `mov` and before the `cmp`. If the ISR makes use of the `eax` register for its operations, it has to push it, and a stack corruption may permit to modify the content of `eax` with the desired label or even with the binary target of the attack. This is not unlikely: `eax` is a general-purpose register which may be used by the ISR.

An additional defense based on binary instrumentation is *Control-Flow Locking (CFL)* [8], which consists in inserting “lock” code before indirect transfer instructions and “unlock” code at each of their valid target. The lock code sets a lock variable to a value, while the unlock code, before proceeding with the execution, verifies whether the value is the lock one. The lock code also verifies if the just-executed code was unlocked and thus allowed to run, otherwise it notifies a violation. The two codes are specular:

```
L_lock:  cmp lck, 0
         jne violation
         mov lck, key
         ret
         ...
L_unlock : call <function>
         cmp lck, key
         jne violation
         mov lck, 0
```

As in the label-based approach, problems may stem by the fact that the flags are possibly altered during the execution of a vulnerable ISR, so the violation can be skipped. In addition, even if the author claims that value of `lck` is stored in a protected memory, it is likely that the one seen above is not the real set of added instructions, and that `lck` is first transferred into a register. This is definitely true when this solution is to be implemented in a RISC machine, where comparisons with memory locations are not allowed. In such a situation, if an interrupt arrives during the manipulation of the lock value, and the register used is pushed because the ISR needs it, then it is possible that it may be restored as corrupted, with consequent defeat of the defense.

C. Hardware-assisted CFI

CFI solutions based on code instrumentation lacks sufficient isolation of the variables and data structures that provide security, as we have shown, as well as they suffer of large overhead. Hardware-based CFI solutions try to overcome these limits. The respect of the CFG is checked at runtime via a *hardware monitor*, which in most cases is directly inserted

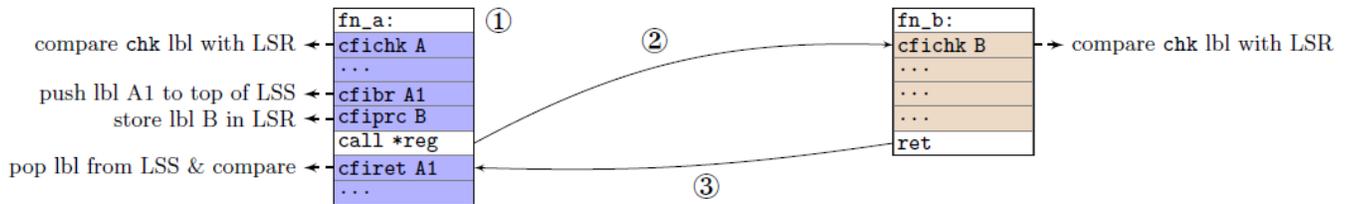


Fig. 2. Tracking of CFG in Sullivan *et al.* solution [42].

into the processor’s pipeline stages [23] [42] [19] [49] but can be also attached externally to the debug interface [30].

Sullivan *et al.* [42] presented a solution in which they modified the soft processor SPARC LEON3 introducing two security-based private data structures, the “Label State Register” (LSR) and the “Label State Stack” (LSS), and five additional instructions to control them. When an indirect call is to be performed, **cfibr** lbl is executed first, which pushes a unique label on top of LSS, indicating the call site. **cfiprc** lbl then saves into the LSR a unique label for the function location. The first instruction of each function is always **cfichk** lbl, which verifies that the content of LSR is lbl. At return time, the processor enters in a state which only accepts **cfiret** lbl, otherwise it is blocked. So all the instructions following an indirect call are **cfiret** in such an architecture. Indirect jumps are instead instrumented with **cfiprj** lbl, which store lbl in the LSR, and at all possible destinations, **cfichk** lbl verifies that the content of LSR is in fact lbl.

However, the hardware implementation of the CFI enforcement does not make it immune to possible breaks due to interrupts. Referring to Figure 2, let us assume that an interrupt is triggered after **cfibr** A1. The ISR reached cannot obviously verify the caller identity with a **cfichk**, because it accepts a static label, unknown at code writing time. The ISR may be vulnerable, and the return address may be tampered with, and again the return site cannot be instrumented. The attacker can thus enter into the middle of any function body, and execute any wondered piece of code. At a certain time, when the **ret** is executed, the top of the LSS is written with A1, so the function returns back to the original call site, and no violation is sensed, as the **cfiret** performs a valid check.

IV. CONCLUSIONS

The present paper analysed the threat of Code Reuse Attacks (CRA) and some of its countermeasures based on complying with the Control-Flow Graph (CFG) with the unpredictability of hardware interrupts. These, in fact, naturally conflict with the static nature of any pre-execution instrument and can open breaches in the defences proposed so far, independently of their actual implementation in software or in hardware. Although exploiting these vulnerabilities to successfully carry out an attack is not trivial, their presence is evident, and it is therefore advisable *to try to lock the door better before someone could learn how to open it*. Nobody can exclude

that somewhere, hidden within the code, there is a sequence of even few but very dangerous instructions, such as setting a password or a key with a default value, or overwriting important memory areas, or that may be exploited to activate additional vulnerabilities to exploit later.

Fine-grained CFI solutions remain today the only effective way to defend against this type of attacks. However, especially when interrupts are frequent, such as microcontroller applications in embedded systems, this particular weakness cannot be ignored, and additional solutions must be adopted.

V. ACKNOWLEDGMENTS

This paper is supported in part by European Union’s Horizon 2020 research and innovation programme under grant agreement No. 830892, project SPARTA.

REFERENCES

- [1] The advanced return-into-lib(c) exploits: PaX case study. <http://www.phrack.org/archives/issues/58/4.txt>, 2001. [Online; accessed 17-June-2019].
- [2] TIOBE Index of May 2019. <https://www.tiobe.com/tiobe-index/>, 2019. [Online; accessed 08-June-2019].
- [3] M. Abadi, M. Budiu, Ú. Erlingsson, and J. Ligatti. Control-flow integrity. In *Proceedings of the 12th ACM conference on Computer and communications security*, pages 340–353. ACM, 2005.
- [4] J. Afek and A. Sharabani. Dangling pointer: Smashing the pointer for fun and profit, 2007.
- [5] S. Alexander. Defeating compiler-level buffer overflow protection. *The USENIX Magazine; login*, 2005.
- [6] F. E. Allen. Control flow analysis. In *ACM Sigplan Notices*, volume 5, pages 1–19. ACM, 1970.
- [7] S. Bhatkar, D. DuVarney C, and R. Sekar. Address obfuscation: An efficient approach to combat a broad range of memory error exploits. In *USENIX Security Symposium*, volume 12, pages 291–301, 2003.
- [8] T. Bletsch, X. Jiang, and V. Freeh. Mitigating code-reuse attacks with control-flow locking. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 353–362. ACM, 2011.
- [9] T. Bletsch, X. Jiang, V. W. Freeh, and Z. Liang. Jump-oriented programming: a new class of code-reuse attack. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pages 30–40. ACM, 2011.
- [10] C. Bresch, D. Hély, A. Papadimitriou, A. Michelet-Gignoux, L. Amato, and T. Meyer. Stack redundancy to thwart return oriented programming in embedded systems. *IEEE Embedded Systems Letters*, 10(3):87–90, Sep. 2018.
- [11] C. Bresch, A. Michelet, L. Amato, T. Meyer, and D. Hely. A red team blue team approach towards a secure processor design with hardware shadow stack. In *2017 IEEE 2nd International Verification and Security Workshop (IVSW)*, pages 57–62, July 2017.
- [12] E. Buchanan, R. Roemer, H. Shacham, and S. Savage. When good instructions go bad: Generalizing return-oriented programming to risc. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 27–38. ACM, 2008.

- [13] N. Carlini and D. Wagner. Rop is still dangerous: Breaking modern defenses. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 385–399, 2014.
- [14] S. Checkoway, L. Davi, A. Dmitrienko, A.R. Sadeghi, H. Shacham, and M. Winandy. Return-oriented programming without returns. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 559–572. ACM, 2010.
- [15] S. Checkoway, A. J. Feldman, B. Kantor, J.A. Halderman, E. W. Felten, and H. Shacham. Can dres provide long-lasting security? the case of return-oriented programming and the avc advantage. *EVT/WOTE*, 2009, 2009.
- [16] P. Chen, H. Xiao, X. Shen, X. Yin, B. Mao, and L. Xie. Drop: Detecting return-oriented programming malicious code. In A. Prakash and I. Sen Gupta, editors, *Information Systems Security*, pages 163–177, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
- [17] P. Chen, X. Xing, B. Mao, L. Xie, X. Shen, and X. Yin. Automatic construction of jump-oriented programming shellcode (on the x86). In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pages 20–29. ACM, 2011.
- [18] Y. Cheng, Z. Zhou, Y. Miao, X. Ding, and R. H. Deng. Ropecker: A generic and practical approach for defending against rop attack. 2014.
- [19] N. Christoulakis, G. Christou, E. Athanasopoulos, and S. Ioannidis. Hcfi: Hardware-enforced control-flow integrity. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 38–49. ACM, 2016.
- [20] C. Cowan, C. Pu, D. Maier, J. Walpole, P. Bakke, S. Beattie, A. Grier, P. Wagle, Q. Zhang, , and H. Hinton. Stackguard: Automatic adaptive detection and prevention of buffer-overflow attacks. 98:5–5, 01 1998.
- [21] S. Das, W. Zhang, and Y. Liu. A fine-grained control flow integrity approach against runtime memory attacks for embedded systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(11):3193–3207, Nov 2016.
- [22] L. Davi, A. Dmitrienko, M. Egele, T. Fischer, T. Holz, R. Hund, S. Nürnberger, and A.R. Sadeghi. Mocfi: A framework to mitigate control-flow attacks on smartphones. In *NDSS*, volume 26, pages 27–40, 2012.
- [23] L. Davi, M. Hanreich, D. Paul, A.R. Sadeghi, P. Koeberl, D. Sullivan, O. Arias, and Y. Jin. Hafix: hardware-assisted flow integrity extension. In *Proceedings of the 52nd Annual Design Automation Conference*, page 74. ACM, 2015.
- [24] L. Davi, A. Sadeghi, D. Lehmann, and F. Monrose. Stitching the gadgets: On the ineffectiveness of coarse-grained control-flow integrity protection. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 401–416, 2014.
- [25] A. Francillon and C. Castelluccia. Code injection attacks on harvard-architecture devices. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 15–26. ACM, 2008.
- [26] Aurélien Francillon, Daniele Perito, and Claude Castelluccia. Defending embedded systems against control flow attacks. In *Proceedings of the first ACM workshop on Secure execution of untrusted code*, pages 19–26. ACM, 2009.
- [27] E. Göktas, E. Athanasopoulos, H. Bos, and G. Portokalidis. Out of control: Overcoming control-flow integrity. In *2014 IEEE Symposium on Security and Privacy*, pages 575–589, May 2014.
- [28] E. Göktas, E. Athanasopoulos, M. Polychronakis, H. Bos, and G. Portokalidis. Size does matter: Why using gadget-chain length to prevent code-reuse attacks is hard. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 417–432, 2014.
- [29] Y. Guo, L. Chen, and G. Shi. Function-oriented programming: A new class of code reuse attack in c applications. In *2018 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9, May 2018.
- [30] Z. Guo, R. Bhakta, and I. G. Harris. Control-flow checking for intrusion detection via a real-time debug interface. In *2014 International Conference on Smart Computing Workshops*, pages 87–92, Nov 2014.
- [31] T. Kornau et al. *Return oriented programming for the ARM architecture*. PhD thesis, Master’s thesis, Ruhr-Universität Bochum, 2010.
- [32] K. S. Kumar and D. Malathi. A novel method to find time complexity of an algorithm by using control flow graph. In *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*, pages 66–68, April 2017.
- [33] A. One. Smashing the stack for fun and profit. *Phrack magazine*, 7(49):14–16, 1996.
- [34] V. Pappas, M. Polychronakis, and A. D. Keromytis. Transparent rop exploit mitigation using indirect branch tracing. In *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*, pages 447–462, 2013.
- [35] P. Philippaerts, Y. Younan, S. Muylle, F. Piessens, S. Lachmund, and T. Walter. Cpm: Masking code pointers to prevent code injection attacks. *ACM Transactions on Information and System Security (TISSEC)*, 16(1):1, 2013.
- [36] J. Pincus and B. Baker. Beyond stack smashing: recent advances in exploiting buffer overruns. *IEEE Security Privacy*, 2(4):20–27, July 2004.
- [37] R. Roemer, E. Buchanan, H. Shacham, and S. Savage. Return-oriented programming: Systems, languages, and applications. *ACM Transactions on Information and System Security (TISSEC)*, 15(1):2, 2012.
- [38] G. F. Roglia, L. Martignoni, R. Paleari, and D. Bruschi. Surgically returning to randomized lib(c). In *2009 Annual Computer Security Applications Conference*, pages 60–69, Dec 2009.
- [39] AliAkbar Sadeghi, Salman Niksefat, and Maryam Rostampour. Pure-call oriented programming (pcop): chaining the gadgets using call instructions. *Journal of Computer Virology and Hacking Techniques*, 14(2):139–156, May 2018.
- [40] F. Schuster, T. Tendyck, C. Liebchen, L. Davi, A. Sadeghi, and T. Holz. Counterfeit object-oriented programming: On the difficulty of preventing code reuse attacks in c++ applications. In *2015 IEEE Symposium on Security and Privacy*, pages 745–762, May 2015.
- [41] H. Shacham et al. The geometry of innocent flesh on the bone: return-into-libc without function calls (on the x86). In *ACM conference on Computer and communications security*, pages 552–561. New York,, 2007.
- [42] D. Sullivan, O. Arias, L. Davi, P. Larsen, A. Sadeghi, and Y. Jin. Strategy without tactics: Policy-agnostic hardware-enhanced control-flow integrity. In *2016 53rd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pages 1–6, June 2016.
- [43] Microsoft Support. A detailed description of the Data Execution Prevention (DEP). <https://support.microsoft.com/en-us/help/875352/a-detailed-description-of-the-data-execution-prevention-dep-feature-in>. [Online; accessed 18-June-2019].
- [44] L. Szekeres, M. Payer, T. Wei, and D. Song. Sok: Eternal war in memory. In *2013 IEEE Symposium on Security and Privacy*, pages 48–62, May 2013.
- [45] PaX Team. PaX Non-Executable Pages Design and Implementation. <https://pax.grsecurity.net/docs/noexec.txt>, 2003. [Online; accessed 17-June-2019].
- [46] M. Tran, M. Etheridge, T. Bletsch, X. Jiang, V. Freeh, and P. Ning. On the expressiveness of return-into-libc attacks. In *International Workshop on Recent Advances in Intrusion Detection*, pages 121–141. Springer, 2011.
- [47] Tzi-Cker Chiueh and Fu-Hau Hsu. Rad: a compile-time solution to buffer overflow attacks. In *Proceedings 21st International Conference on Distributed Computing Systems*, pages 409–417, April 2001.
- [48] Yubin Xia, Yutao Liu, H. Chen, and B. Zang. Cfimon: Detecting violation of control flow integrity using performance counters. In *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012)*, pages 1–12, June 2012.
- [49] J. Zhang, B. Qi, Z. Qin, and G. Qu. Hcic: Hardware-assisted control-flow integrity checking. *IEEE Internet of Things Journal*, 6(1):458–471, Feb 2019.
- [50] Mingwei Zhang and R Sekar. Control flow integrity for cots binaries. In *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*, pages 337–352, 2013.

From Abstract Modeling of ADAS Applications to an Accelerator-based Hardware Realization

Samira Ahmadi Farsani, Katayoon Basharkhah, Amin Mohaghegh, Zainalabedin Navabi
School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
{samira.ahmadi.fa, basharkhah.kt96, amin.mohaghegh, navabi}@ut.ac.ir

Abstract— Abstraction alongside with multi-level modeling of embedded real-time systems, such as ADAS applications, has gained attention for system analysis and hardware realization. Due to the complexity of the state of art embedded digital signal processing, pre-evaluation before design can reduce time and performance hinders. As a system-level language and framework, SystemC is used in this work to exercise its suitability for hardware design space exploration (DSE). The methodology used in this paper provides a hardware/software, multi-level unified modeling environment and shows hardware realization of a SystemC model considering concurrency of the modules, modularity, and synchronization features. A vehicle detection application is evaluated as a case study.

Keywords— SystemC, design space exploration, ADAS, vehicle detection

I. INTRODUCTION

Today's trend for embedded systems is going toward implementing system on chips (SOC) including multiple processors, high-speed communication interfaces and reconfigurable processing elements. With the increased complexity of such digital circuits, the need for higher performance, much more productivity, time to market and lower cost increases in turn. These factors made the SOC design a challenging topic.

Hardware, analog and software co-simulation can result in more optimal platforms by eliminating the need for redesigning. Furthermore, traditional RTL based designs impose long-time simulation. This can be a major drawback during test and verification since based on Moore's law, the amount of test vectors in a verification rises by a factor of 100 every six years, which is 10 times the increase of the number of gates on a chip [1]. To overcome these problems, there is a need for tools that work at higher levels of abstraction and speed up the simulation.

Modern cars include technology to expand vehicle safety and more generally road security. This concept is known as an advanced driver assistant system (ADAS) that has received considerable attention in recent decades. ADAS systems act by some methods to alert the driver or by taking over control of the vehicle. Computer vision with a combination of both radio detection and ranging (RADAR) and light detection and ranging (LIDAR), is at the forefront of technologies that enable the evolution of ADAS [2]. The implementation of the computer vision-based ADAS applications in a real automotive environment is not an easy procedure. Embedded vision systems for driver assistance need to set a trade-off between several requirements such as dependability, real-time performance, low power consumption, fast time-to-market and etc. [3]. On the other hand, the increase of sensor processing data and the complexity of the algorithms seek more powerful processing platforms. Heterogeneous

platforms composed of microcontrollers and hardware accelerators are at the bleeding edge of ADAS implementations [3].

SystemC, a class of C++ language, is a good option for electronic system level of abstraction. This standard language is able to create every model with any complexity using specified wrappers, representing functional elements of digital systems and can be used to verify and simulate functionality and behavior of such systems faster. SystemC can be used for evaluating the architectural and functional trade-offs and selecting the best architecture and procedure through design space exploration. So DSE is facilitated in modern design with using SystemC.

This paper presents a design flow that takes an algorithmic software description of a vehicle detection as an ADAS application, and steps through the design stages for creating a set of communicating hardware modules that implement the starting algorithm. The application is described in C/C++, and the final hardware representation is in SystemC.

The next section presents a brief description of SystemC indicating how it will be used in high-level system design. The section that follows discusses vehicle detection and an algorithm for its implementation. Steps for extracting SystemC modules and their corresponding communications are discussed in Section IV. Section V shows simulation results of the hardware realization of the vehicle detection application that is now represented in SystemC.

II. DESIGN WITH SYSTEMC

SystemC is a high-level modeling language that uses a discrete event simulation kernel based on C++. Different scenarios of an embedded system can be verified using SystemC modeling. This provides a platform for design space exploration. SystemC can pick and choose levels of abstraction from simple gate level to high-level abstracted models.

SystemC consists of several SC_MODULES, instantiated inside the main module. The modules are connected through their ports using channels. An *sc_signal* is a built-in SystemC channel for RTL descriptions, while other channels for higher abstraction are also offered by SystemC. Definition of other custom channels are also possible in SystemC that can be used for special types of data and handshaking mechanisms.

A set of *transport* channels are also defined in SystemC that mainly focus on memory-mapped IO and memory communications. This is referred to as TLM-2.0, which is now widely used in embedded applications.

There exist more complex channels like FIFO, stack and even caches. This is done by registering concurrency of processes inside constructor's SC_THREAD anywhere from

gate level to behavioral software program describing the system.

III. VEHICLE DETECTION IN ADAS

Determining the position of the vehicles in front of one's own is key information to help driver assistance systems to increase a driver's safety and accident prevention. So a major function of ADAS is vehicle detection using computer vision technologies [4]. There are several methods and techniques to implement vehicle tracking and detection systems.

Different popular techniques of vehicle detection are discussed in [2]. Reference [4] gives an accurate identification and high-performance results compared to other existing methods. Work described in this reference achieves an accuracy of 95.8% at 30 frames per second for detecting proceeding vehicles tested on highways in the daytime.

As a case study, we select the system of [4] with some modifications and implement its detection mode. In this system, the features of the vehicle in front are extracted and recognized by the following refined image processing algorithm. The flow of this algorithm is shown in Fig. 1.

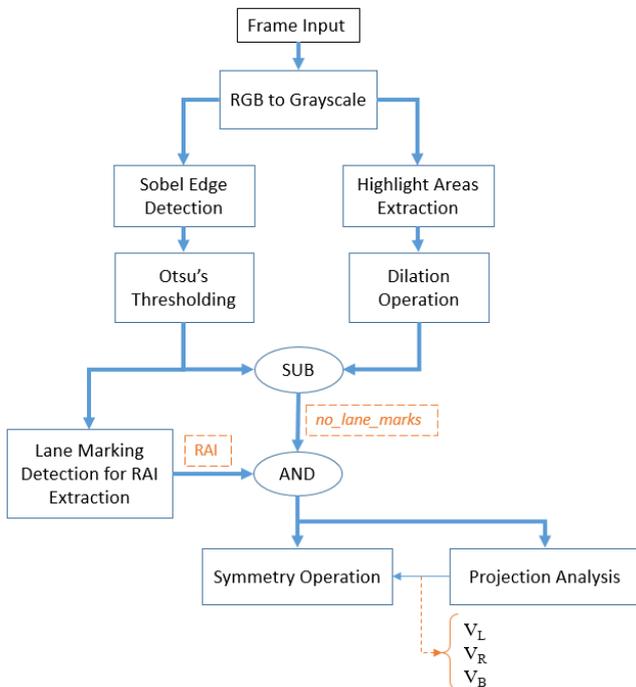


Fig. 1. Flow of vehicle detection algorithm

At the beginning, to reduce the computing complexity, the color space of the image will be transformed from RGB to grayscale.

For removing the interference and strengthening the profiles of vehicles, the lane marks of the image are removed to generate *no_lane_marks* image. Then road area image (RAI) is extracted. The *no_lane_marks* image along with RAI is used to execute the logical-AND operation, yielding a footprint image. The left, right and bottom boundaries of the vehicle block are obtained by analyzing the footprint image. The footprint image may include wet spots, the seam of a bridge, and shadows of passing vehicles. An operation, that is referred to as the symmetry operation, determines whether these footprints are related to vehicles [4].

IV. SYSTEMC IMPLEMENTATION

Software implementation for most of the computer vision algorithms have extensively been researched, and are available. Implementation of these algorithms on a hardware platform is necessary, and is more efficient than the software running on a processor. This becomes more important for real time applications such as ADAS. For a more efficient hardware implementation, and one that we can start with the software specification of the ADAS application, and in order to be able to examine hardware and software implementations in the same environment, we use SystemC and its C++ based library.

In the first step, we write the C++ code of the algorithm using the OpenCV library. For this, corresponding C++ modules have been developed based on the block diagram in Fig. 1. Then SystemC is used to contain individual and independent modules that can be executed concurrently. This will define a system of independent accelerators each of which perform a given task related to the blocks of Figure 1. Then, based on the data they use, concurrent execution of these blocks will be decided. Depending on the type and data requirement of each of the accelerators (concurrent SystemC modules), proper communication links (SystemC channels) will be defined.

A. Communications

Various SystemC modules implementing the vehicle detection algorithm communicate via buffers and memory blocks.

In OpenCV computations, image inputs are considered as 2D arrays (*Mat* type for image). To realize the hardware correspondence, instead of 2D arrays a memory must be used to save the images. Therefore, the frame input after color space conversion is stored in memory as a vector of elements in row-major order.

Implementation of the memory (with the size of rows \times columns of the image) in SystemC is shown in Fig. 2. Three interfaces define how this memory is used. The first interface is *requestMem* that is called when an initiator needs the attention of the memory. The *memForward* interface takes back the requested message (read or write) to the memory, and the third interface, *memBackward*, satisfies the request. These interfaces are implemented in what SystemC refers to as channels. The channel used here is *memoryAccess* channel that is a custom channel that we have developed for memory accesses in this ADAS application.

The emphasis of this work is mainly on the computations performed by SystemC modules and their concurrencies, and not as much on their communications. Because of this, we have only developed the *memoryAccess* channel and use it for most of the communications of the SystemC modules. The next step of this work would be to deviate from a uniform form of communication and design channels that are custom made for the specific communication requirements between communicating SystemC modules. This will reduce hardware requirements for interfacing modules, and will have the added advantage of minimum hardware used for commutation links.

```

//memory_channel
class mem_req_if {...};
class mem_res_if {...};
class memory_access {...};
void memory_access::get_mem(int addr, short int& data) {...}
void memory_access::mem_forward(int& addr) {...}
void memory_access::mem_backward(short int data) {...}

//memory_module
SC_MODULE(memory)
{
    sc_port< mem_res_if > in;
    int addr;
    short int* data_array;
    ifstream input_file;
    void mem_response();
    SC_CTOR(memory)
    {
        input_file.open("imgfile.txt");
        data_array = new short int[154896];
        SC_THREAD(mem_response);
    }
};
void memory::mem_response()
{
    for(int i = 0; i < 154896; i++)
    {
        input_file >> data_array[i];
    }
    cout << "memory initialization finish" << endl;
    while(1)
    {
        in -> mem_forward(addr);
        wait(1, SC_NS);
        in -> mem_backward(data_array[addr]);
    }
}

```

Fig. 2. SystemC code of the memory

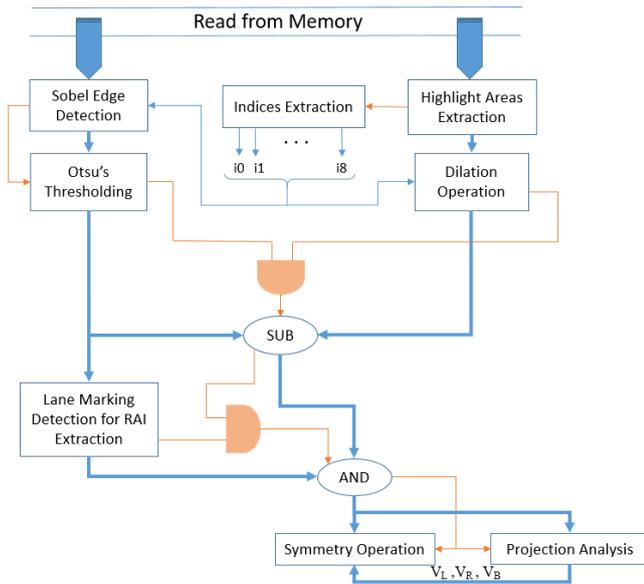


Fig. 3. Block diagram of SystemC model

B. Computations

Computation blocks for the implementation of the algorithm of Fig. 1 become SystemC modules as shown in Fig. 3. The blocks in this block diagram generally correspond to C++ codes of the blocks of the flowchart of Fig. 1, with a mechanism added for synchronization between these modules. Each module issues a *done* signal (orange-colored signals in Fig. 3) after completing its process in order to synchronize it with its preceding modules.

The block diagram of Fig. 3 also includes an *Indices Extraction* module and two AND blocks. The indices extraction is used for the generation of proper memory references by the individual modules. The AND functions

have been included inside SystemC modules, and are used for inter-module handshaking and synchronization.

SystemC modules describing the computation units begin with proper *sc_ports* for signal communication with the other modules and for channel communications. The SC_MODULE of a given computation includes a constructor (SC_CTOR) within which an SC_THREAD registers the function of the computation as a concurrent process. The SC_THREAD also defines the activation of the function making it sensitive to handshaking signals that invoke the concurrent process.

Within these functions, an invoked C++ program describes the functionality of the computation block. Often, the original C++ program of the software implementation (i.e., Fig 1) is the starting point for the SystemC implementation. The C++ functions of *Sobel Edge Detection*, *Otsu's Thresholding*, *Highlight Areas Extraction*, *Projection Analysis*, and *Symmetry Operation* have been written behaviorally to process 8-bit pixels. AND and SUB modules are implemented with logical-AND and subtraction functions, respectively. The next several sub-sections present more details of the SystemC descriptions of the computation units and their concurrent registered functions.

1) *Highlight Areas Extraction*: The *highlight* SC_MODULE generates a binary image from the input image file. The road surface is sampled at five fixed areas. Average of these samples are used as a threshold for binarizing each pixel. This module is partially shown in Fig. 4.

```

void highlight::do_highlight()
{
    highlight_complete.write(false);
    wait(5, SC_NS);
    for(int i = 0; i < imagesize; i++)
    {
        img_mem_bus -> get_mem(i, imgmem); //reading from memory
        wait(1, SC_NS);
        // comparison with threshold for binarizing each pixel
        if(imgmem > mean)
            highlight_out[i].write(255);
        else
            highlight_out[i].write(0);
        wait(1, SC_NS);
    }
    highlight_complete.write(true);
}

```

Fig. 4. SystemC code of the *Highlight Areas Extraction*

2) *Indices Extraction*: The *Indices Extraction* module (*indexing*) extracts the indices of 9 pixels in each iteration of Sobel filter and dilation operation using the algorithm in [5] shown in Fig. 5. Inputs *rows* and *cols* represent the row number and the column number of the image, respectively. This module is sensitive to the positive edge of the *done* signal of the highlight areas extraction module. This way, indices extraction will be started right after highlight areas extraction is completed.

The for-loop in this module can easily be translated to a hardware accelerator. Fig.6 shows hardware implementation of this module for which a hardware accelerator could be used. Similar accelerator-based implementations corresponding to the for-loops for other modules also exist. For illustrating the concept of hardware translation of the for-loops, we have used the *indexing* module that is simple and can easily be described

in a simple diagram. Hardware correspondence of other models that follow will not be shown.

```

SC_MODULE(indexing) {
    sc_in< int > cols;
    sc_in< int > rows;
    sc_in< bool > highlight_complete;
    sc_out< int > i0,i1,i2,i3,i4,i5,i6,i7,i8;
    int k;
    void do_indexing()
    {
        wait(2, SC_NS);
        for (int i = 1; i < (rows - 1); i++) {
            k = i * cols;
            for(int j = 1; j < (cols - 1); j++) {
                i0 = k - cols + j -1;
                i1 = k - cols + j;
                i2 = k - cols + j +1;
                i3 = k + j - 1;
                i4 = k + j;
                i5 = k + j + 1;
                i6 = k + cols + j -1;
                i7 = k + cols + j;
                i8 = k + cols + j + 1;
                wait(1, SC_NS);
            }
        }
    }
};
SC_CTOR(indexing){...}

```

Fig. 5. SystemC code of the *Indices Extraction*

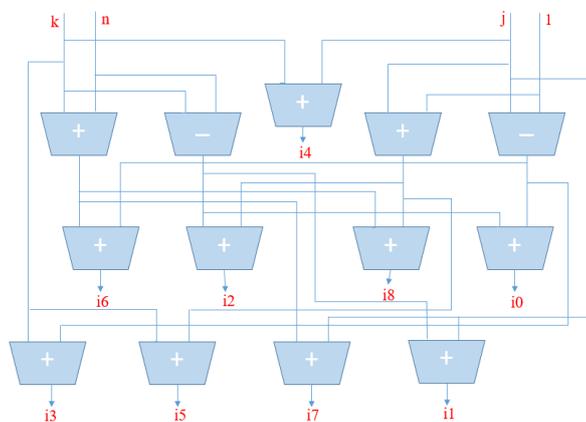


Fig. 6. Hardware of *Indices Extraction*

3) *Sobel Edge Detection*: A Sobel filter is used for vehicle edge detection. In the Sobel filter, a 3×3 sub-window of the image is convolved with Sobel masks in each iteration. On the other hand, for the convolution process, the image is scanned from left to right and top to bottom. The indices of the 9 pixels for the convolution process are prepared by indices extraction module in each iteration. The SystemC thread of *SobelFilter* is sensitive to the calculation of the last index of the *indexing* module. This module issues *sobel_finish* for the next module to start its thread. SC_MODULE of the module is shown in Fig. 7.

```

SC_MODULE(SobelFilter)
{
    sc_in< int > i0,i1,i2,i3,i4,i5,i6,i7,i8;
    sc_port< mem_req_if > img_data;
    sc_in<int> rows;
    sc_in<int> cols;
    sc_in<bool> sobel_finish;
    sc_out< short int > edgeram[153306];
    int sum;
    int xgrd;
    int ygrd;
    int i;

    void do_Sobel(){...}

    SC_CTOR(SobelFilter){...}
};

```

Fig. 7. SystemC code of the *Sobel Edge Detection*

4) *Otsu's Thresholding*: In *Otsu's Thresholding*, after calculating a threshold from the histogram of the image each pixel is binarized. Otsu algorithm performs a variance analysis processing to find the optimal threshold. This module is sensitive to the positive edge of the *done* signal of the Sobel filter, i.e., *sobel_finish*.

5) *Dilation Operation*: Dilation operation is applied to intensify the binary road image. The image is convoluted with a structural element like the Sobel filter. The center point of the structural element in the image is replaced with the OR operation of the pixels in the structural element [6]. A 3×3 square is selected as the structural element to use the outputs of *Indices Extraction* module (i.e., *indexing*) and reduces the computations.

As such, this module is sensitive to the calculation of the last index of the *indexing* module. As described above, the *SobelFilter* module is also sensitive to the last index calculation. As shown in Fig. 3, the *SobelFilter* and *dilation* modules that start after *indexing* are concurrent modules that are assigned to concurrent accelerators in the implementation of hardware.

An important observation here has to do with the use of *indexing* that is required for both *SobelFilter* and *dilation* modules. These two modules perform each of their iterations when nine indexes are prepared by *indexing* module. The indexes are prepared one at the time, and after the ninth index, their next iterations begin. When all indices of memory elements are completed, *SobelFilter* and *dilation* will be terminated.

On the other hand, as shown in Fig. 3, the *dilation* module requires termination of *Highlight Areas Extraction* (*highlight*), that the *SobelFilter* module does not. This means that *SobelFilter* could start sooner at the same time as the *highlight* module if the *indexing* module could be duplicated, one to provide indexes for *SobelFilter* and one to wait for the completion of *highlight* and then provide indexes for *dilation*. This is a clear case of hardware choices that we explore during DSE when a SystemC environment is being used. The hardware and function of *dilation* module are shown in Fig. 8.

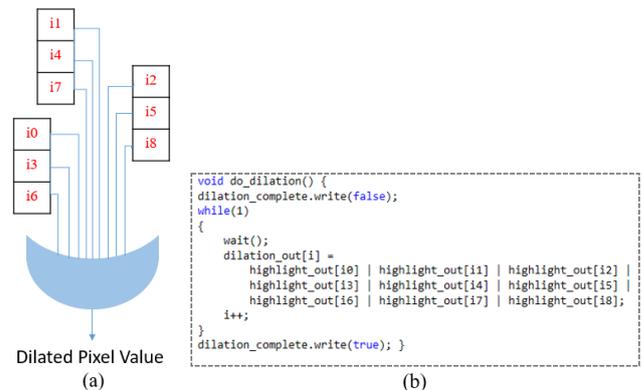


Fig. 8. (a) Hardware, (b) Function code of *Dilation Operation*

6) *RAI Extraction*: For extraction of the road area image (RAI), firstly a Hough line detector is applied to extract the position of the lane mark pixels in the image. Then the part of the image between the two extracted lane marks is saved

as RAI. In C++ implementation of RAI extraction, OpenCV functions such as HoughlinesP (Probabilistic Hough Transform), Line () and Point () have simplified the process as shown in Fig. 9.

On the other hand, writing the behavioral code of these functions directly in C++ to process a linear array and not a 2D array (*Mat* type for an image in OpenCV) is an inefficient and complex work. Because of this complexity, in our first attempt to complete the SystemC implementation of vehicle detection, we simply use the RAI extraction module of OpenCV library and bracket it with actual signals in an SC_MODULE. This means that our SystemC module for this looks exactly like a hardware module with proper signals and channels. This module also has a concurrent process that is registered as an SC_THREAD. The registered function uses available OpenCV utilities for implementing its functionality.

Using OpenCV functions require input and output image types, e.g., *Mat* 2D type. The registered thread in our SystemC module is responsible for reading our linear memory and turning it into an image, and when the OpenCV function has completed, the image is turned back into our hardware oriented linear memory format.

Had there been a hardware implementation for the OpenCV code, our using this implementation would imply that we were using an existing IP Core instead of actually designing its hardware ourselves. In this case, using the SystemC SC_MODULE bracketing would imply the generation of a hardware wrapper to adapt the IP Core to the other modules of Fig. 2.

```

// Create a vector to store lines of the image
vector<Vec4i> lines;

// Apply Hough Transform
HoughLinesP(inputimg, lines, 1, CV_PI/180, 210, 0, 0);

// Draw lines on the image
for (size_t i=0; i<lines.size(); i++) {
    Vec4i l = lines[i];
    line(dst, Point(1[0], 1[1]), Point(1[2], 1[3]), Scalar(255, 0, 0), 3, LINE_AA);
}

```

Fig. 9. OpenCV code of line detection

7) *Projection Analysis*: In the projection analysis block, the left, right and bottom boundaries of the vehicle are obtained and denoted as V_L , V_R , and V_B respectively. So the vehicle image block is defined with height and width. The width and height of the vehicle image block is $W = |V_R - V_L|$ and $H = 0.8W$, respectively.

8) *Symmetry Operation*: Symmetry operation module finds the most symmetric axis by minimizing the symmetry measure $S(j)$ where j is the position of the symmetry in the vehicle image block [7]:

$$S(j) = \sum_{i=V_B-H+1}^{V_B} \sum_{\Delta x=1}^{W/2} |P(i, j + \Delta x) - P(i, j - \Delta x)|$$

For $V_L < j < V_L + W$, $j_{sym} = \operatorname{argmin} S(j)$ and $\min S(j) < S_{th}$

$P(i, j)$ denotes a component in the vehicle image block. As it can be seen from the above expression, the symmetry operation can be completed by three nested loops. As with the other hardware blocks, the *symmetry* module can also be

realized with an accelerator that accesses the vehicle part of the input image memory.

V. RESULTS

The hardware implementation that we presented above begins with reading a linear memory. This memory is initially filled with an image from an external file. The output arrays of *Otsu's Thresholding*, *Dilation Operation*, *SUB*, *RAI Extraction* and the *AND* modules in SystemC also become available in blocks of memory. These outputs are then stored in five separate files for the input image.

In order to verify operation of the intermediate steps, the file outputs, that reflect contents of the linear memory used by various modules, are written into external files for display. The size of the input image is 336 (rows) \times 461 (columns) pixels which becomes the image size of all hardware modules.

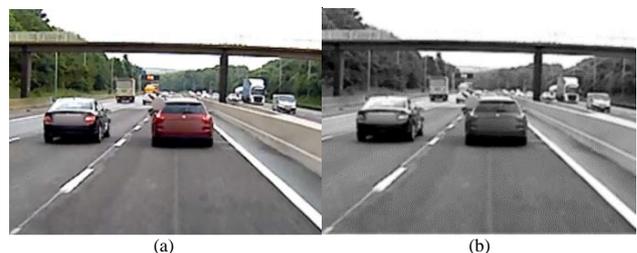


Fig. 10. (a) Input image. (b) Grayscale image

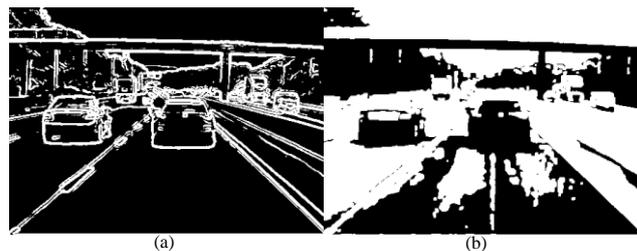


Fig. 11. (a) Otsu's thresholding output. (b) Dilation output.

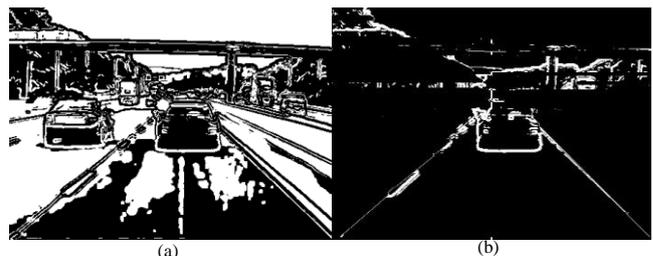


Fig. 12. (a) Subtraction output. (b) RAI extraction output.

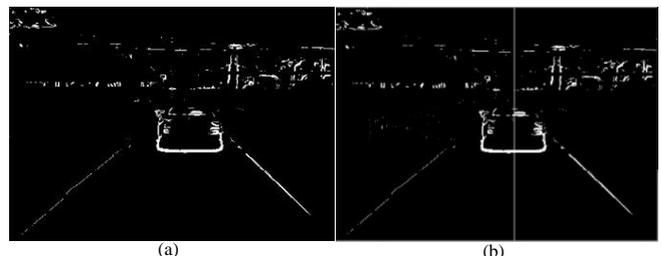


Fig. 13. (a) AND module output. (b) Calculated symmetry axis of vehicle block.

The RGB and grayscale input images are shown in Fig. 10. The grayscale image first is passed through the Sobel filter that generates an edge-detected image. Then it is binarized

with Otsu's thresholding as shown in Fig. 11(a). The highlighted image, yielded by highlight areas extraction module are dilated using the dilation operation. The result of this block is shown in Fig. 11(b). The *no_lane_marks* image, the output of the subtraction module, and the road area image (RAI) are shown in Fig. 12. The footprint image is achieved by logical-AND of two images, RAI, and *no_lane_marks* shown in Fig. 13(a). Fig. 13(b) shows the vehicle block image with the calculated symmetry axis.

VI. CONCLUSION

This paper has presented a design flow that takes an algorithmic software description of a vehicle detection as an ADAS application, and steps through the design stages for creating a set of communicating hardware modules that implement the starting algorithm. The application has been described in C/C++, and the final hardware representation is in SystemC. SystemC has been used for evaluating the architectural and functional trade-offs and selecting the best architecture and procedure through design space exploration.

REFERENCES

- [1] Moore, G. E. 1965. "Cramming more components onto integrated circuits". *Electronic Magazine*, Vol. 38, No. 8, 19 April 1965. Retrieve on 11 November 2006
- [2] Sowmya Shree, and A. Karthikeyan. "Computer vision based advanced driver assistance system algorithms with optimization techniques-a review," *Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 821-829. IEEE, 2018.
- [3] Gorka Velez and Oihana Otaegui. "Embedding vision-based advanced driver assistance systems: a survey." *IET Intelligent Transport Systems* 11, no. 3, pp. 103-112, 2016.
- [4] Ying-Che Kue, Neng-Sheng Pai and Yen-Feng Li. "Vision-based vehicle detection for a driver assistance system." *Computers & Mathematics with Applications* 61, no. 8, 2011.
- [5] Nazma Nausheen, Ayan Seal, Pritee Khanna, and Santanu Halder. "A FPGA based implementation of Sobel edge detection." *Microprocessors and Microsystems* 56, pp. 84-9, 2018.
- [6] Randika Perera and Swapna Premasiri, "Hardware implementation of essential pre-processing & morphological operations in image processing," *6th National Conference on Technology and Management (NCTM)*, January 2017 .
- [7] H.Y. Chang, C.M. Fu and C.L. Huang, "Real-time vision-based preceding vehicle tracking and recognition," *IEEE Intelligent Vehicles Symposium*, pp.514-519, 2005.

Unified STIL Flow: A Test Pattern Validation Approach for Compressed Scan Designs

Slimane Boutobza

slimane@synopsys.com

Andrea Costa
Test Automation

Synopsys Inc
Montbonnot, France
costa@synopsys.com

Sorin Popa

spopa@synopsys.com

ABSTRACT – With the growing complexity of SoCs (system on Chip), and the explosion of test data volume, test patterns validation is becoming a critical step within the test flow. For modern large designs, detecting most issues at the level of the ATE (Automatic Test Equipment) is no longer a viable solution. Recent approaches rely on dedicated tools and flows prior to tester, to validate test patterns, and reserve ATE to only screening real defect issues on the test-chip. This allows for early detection of successive and cumulative modeling and implementations issues. In [1] and [2] we introduced an approach for efficient test patterns validation. In the present paper, we address the problematic of cost-effective validation effort, with regards to two main factors, namely, test-time and ease of use. We introduce an original technique called “Unified STIL Flow” (USF). It allows for significant simulation/validation time acceleration (from few weeks to few days) through appropriate algorithmic manipulations and data representations. Compared to our original flow (Dual STIL Flow-DSF), it achieves 2X runtime improvement and 2X memory reduction while greatly simplifying it from user point of view. This technique is an industry proven methodology successfully used today by SC companies in their daily test patterns validation.

Keywords – Validation, test, ATPG, STIL

I. INTRODUCTION

The need for high test quality while minimizing test cost has been a constant quest for semiconductor’s industry [3]. This already complex task has been further complicated by the constant need to decrease development cycle time in order to meet aggressive time to market requirements. To add to this, these factors are interrelated and very often antagonist (e.g., increasing test quality, naturally tends to increase test cost and TTM, while reducing test-time will impact fault coverage -test quality). Thus, very often a subtle tradeoff must be found. In general, no concession is made on the test quality, so that the problem to solve becomes “find the lowest test cost for a given/desired test quality”.

While the factors contributing to test cost are diverse and very often interrelated, a careful analysis shows that the main contributors are the complexity of the test flow and the overall test time [4]. The former has an evident relationship (longer test time means longer engineer worktime, tools and licenses exploitation, more resources), while the later has subtler yet a strong one (a complicated flow requires storing various intermediate data using different formats, tools and time to convert between these formats, longer steps with greater chance to introduce human/machine errors). In this paper, we propose a contribution to reduce test cost by essentially

targeting test time reduction and test flow simplification in context of random logic testing.

Historically, the response of EDA companies to such request was to provide various flavor of scan compression solutions (ranging from combinational to sequential compressors, using register-based or BIST-PRPG based techniques) allowing both test data volume (TDV) and test-time reduction. In addition to spatial and temporal compression, they provide a uniform interface to control the IP at the board and chip level (JTAG and 1500 interfaces) and a low-cost autonomous solution (no need for expensive Testers to control the IP). Unfortunately, with the ever-growing SoC sizes and multimillion FFs (nowadays, having 20-40 M FFs design is not rare, especially for graphical and router/networking applications) resulting in up to hundreds of thousands of test patterns, the contribution of such compression solutions remains limited and the test cost problem is far from being solved.

A typical test suite for random logic testing (figure 1) uses a structural DFT insertion to improve controllability and observability of the circuit, followed by an Automatic Test Pattern Generation (ATPG) allowing for clear test quality assessment through a succinct fault coverage metric, and finally, a Tester validation. Prior to ATE testing, patterns validation (test-sign-off) is a critical step to ascertain the correctness of the generated test vectors and reduce their debug cost by allowing an early and flexible debug capability. While DFTIP generation and insertion can take relatively reasonable time (within a day), and ATPG still most of the time affordable (few days at max), ATE testing and patterns validation (Patval) can both easily exceed several weeks of overall validation time. Thus, with the exclusion of the tester step (figure 1) Patval constitutes the critical path in this EDA tools suite, and by the same token, the best opportunity for test time saving. Therefore, in this paper, we specifically address test-time reduction and flow simplification related to the Patval step. The emergence of multimillion gates SoC, with its associated dozen and even hundreds of thousands test patterns, further exacerbated the importance of test pattern validation.

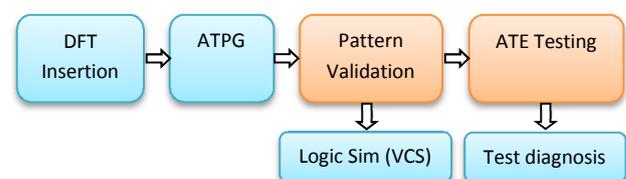


Fig 1: Test Flow

The first obvious solution for test time optimization is automation to reduce the turnaround for TB creation and test vector translations. In [1] we addressed such aspect by presenting an automatic TB generator. It is based on the de facto standard nowadays, the STIL language [5] that allows for compact and optimized descriptions of test vectors, their timing and the expected response. Nonetheless, automation alone is not enough to effectively validate the huge set of test patterns of today's designs. It needs to be reinforced by additional effective strategies.

In fact, the optimizations performed for DFT and ATPG steps are also beneficial to Patval step. A full or partial scan-based testing adopts a regular sequence consisting into 1-setup the DUT, 2- shift-in test vector, 3- capture data, 4- shift-out captured response. Step 1 is done once. A common ATPG technique is to overlap steps 2 and 4, that is, while shifting-in pattern i , we also shift-out pattern $i-1$, which divides the overall pattern application time by 2. Today's legacy scan architecture result in long scan chains of several thousand cells, inducing simulation rate of 1 pattern /day and imposing by the same an upper bound limit for the simulated patterns (at most, few dozens). The DFT-based scan-compression techniques, despite the introduced debug and diagnosis complexity, bring a considerable improvement. They allow reducing by 2-10x the scan chain length (typical 1000 down to 250 cells), hence the corresponding test-time. The cumulative effects of such techniques allow attaining few hundred cycles per test pattern. While such ratio can be affordable for a hardware-based ATE system, it is not for software-based (logic simulator) Patval. The simulation of such test patterns can easily take several weeks if performed in nominal manner, that is in serial mode using the inserted DFT interface. Such serial simulation may be mandatory for modes like timing simulation, eventually with backed annotated SDF, at least for a small subset of test-patterns. But it turns out as a huge waste of time when it comes to pure functional validation of these test patterns.

In fact, besides these techniques, Patval could also bring its own specific optimization. The algorithmic based technique proposed in [6] leverages Patval high flexibility to control and observe internal DUT structures to accelerate the simulation. It's based on the principle of parallel access (force and strobe) to internal scan-cells, which allows reducing the n serial cycles into one single cycle. Parallel simulation is of tremendous importance for test sign-off. We are talking of 10-100X (1 to 2 order of magnitude) acceleration. This technique was implemented in our test tool [7] and by [8] and used for years under DSF (Dual STIL Flow) name. A major drawback of such technique is the impossibility to perform parallel simulation in presence of scan-compression designs and the need for additional generated and reformatted test patterns from the ATPG engine. These patterns are stored in STIL that cannot be consumed by the ATE. They are also redundant with the original STIL file which further complicate the Patval

flow (see section II). Despite these limitations the basic concept is powerful and should not be discarded at once. In this paper we present a novel methodology, called USF that relies on the same principle as of [6] but exploits only its optimal part. USF is a methodology based on the STIL file as unique DB (Unified stands for reducing two STIL modes to use one single STIL file). It exploits all its fields and data to ensure nominal (serial) simulation/validation and mimic the ATE behavior. But in addition, augments the STIL with dedicated compact and compliant STIL structures to derive parallel simulation from the nominal serial STIL targeted for the ATE. The generated USF testbench (TB) uses HDL modules functionally-equivalent to the Compressor Decompressor netlist DFTIPs, with the following properties: a) they are simulation oriented, thus no synthesis style is need, and HDL simulation can be fully exploited, b) they have a behavioral implementation, c) they are targeting parallel simulation, thus all sequential behavior is recoded and accelerated to fit in one single cycle. Besides the parallel acceleration goal, the USF approach ensures other subsidiary goals: 1) parallel acceleration should not come at price of sacrificing coverage of DFT IP (i.e., while testing the DUT, the generated ATPG patterns test also the inserted DFTIP compression logic) 2) a good flexibility that allows for other simulation/validation modes when needed (from the same databases/files) such as NShifts and mixed serial/parallel modes. To be effective, this flexibility must be provided at runtime to overcome time-consuming recompilation, 3) ease of use so that we are not transferring the time saving problem from one place (simulation) to another (settings, maintenance...) by providing a simple and robust flow with as reduced entry points as possible. The USF approach first perform STIL-HDL translation [2] devising a trustworthy and efficient approach to port the problematic from a cycle-based test domain into a simulation domain that mimics the behavior of the tester while ensuring accelerated validation and easier debuggability. Compared to the original DSF parallel simulation flow, it brings a considerable QoR improvement.

In the sequel, we first describe overall requirements and challenges faced by the USF methodology, then give its general concept in section III. Section IV describes STIL compliant augmentations to capture certain DUT properties, while we explain the USF behavior in section V using different simulation modes. We finish the paper by presenting some experimental results and concluding remarks in section VII and VIII respectively.

II. PROBLEM STATEMENT AND CHALLENGES

The general problem of "test time reduction" is a very large topic that involves EDA software, test architecture (scan compression) and resources (available hardware and licenses) and cannot be fully tackled in one single paper. Nevertheless, the technique addressed here is test time saving per test unit (test pattern) and is the main saving factor. Therefore, we are not concerned here with

the concurrent simulation of several pattern sets that is orthogonal and can be coupled with our technique to further improve runtime gain (a classical Operating Research problem to optimize allocation of N pattern sets to M Simulator Licenses).

Historically, the problem of test patterns validation acceleration was tackled using a direct implementation of [6] which resulted in inefficient solutions [7], [8] with limited simulation capabilities. The root cause of such limitation is their need for a dual STIL flow (for scan compression designs). That is, the validation uses a serial STIL file (serial test data) to perform the scan chain test and/or to ensure a full DFT circuit validation, and a parallel STIL file, for medium to big size designs, to bypass the serial load/unload of the internal scan chains thus allowing to accelerate the validation process and coping with the long serial simulation time. Such Dual STIL Flow (DSF) presents indeed a set of limitations:

- Complex and confusing validation flow. We need to maintain and manage two equivalent databases (normal/serial STIL file targeted for ATE test, and parallel STIL targeted for validation) and store, invoke and log the right STIL simulation.
- Adds another complexity layer by providing the user with a (not natural) parallel STIL format that is not obvious to understand (e.g., specific *internal_load_unload* procedure, usage of pseudo scan-in and scan-out...).
- Longer test-time (need for ATPG tool to generate this time-consuming huge data).
- Disk space requirements and time and memory to write out and read in these intermediate multi-GB files.
- Low coverage since it completely bypasses the Comp Decomp testing.
- Low validation performance due to the need to load and simulate this huge DB.
- Failure format not compliant with standard diagnosis.
- Most important of all, in the cases of scan compression designs, the DSF parallel STIL file doesn't allow neither for NShifts nor for mixed serial/parallel simulations (see section V).

To cope with these various drawbacks and provide a best in class validation flow, we established a minimal set of requirements:

- 1- Propose a flow based only on the actual STIL file targeted for the Tester.
- 2- Provide an accelerated validation solution (nominal requirement for this project).
- 3- Ensure at runtime this acceleration (parallel sim) while enabling serial sim.
- 4- Ensure mixed serial/parallel and parallel with n serial shifts (NShifts).
- 5- Ensure at worst, same runtime and memory QoR as the existing DSF solution and same debug capabilities.
- 6- Provide a simple and easy to use overall validation flow.

Such requirements raised several questions and challenges that we needed to address.

a- What to model and which support/language to use? We chose to reuse the STIL despite some weakness, in order to ensure a simple flow with single main entry file. This task is not straightforward, since it requires first, detailed inventory of additional required information (mainly information related to the inserted DFT IP) and then, an efficient modeling of this information in the STIL file while respecting the overall STIL syntax (addressed in section IV).

b- Figure out a technical solution to derive from a unique STIL file, patterns that are valid for different simulation modes. While the existing flow requires a parallel STIL for parallel simulations and a serial STIL for serial simulations, and provides no mixed simulation, the new project aimed to provide all these simulation modes derived from a unique and same (USF) STIL file.

c- Better QoR than DSF in parallel simulation mode. To overcome negative performance impact, we rely on compact and optimized compressed test data and derive parallel data on the fly (at runtime) using cycle-optimized efficient behavioral-based CODECs (section V.B and VII)

d- Extend DSF capabilities: how to efficiently derive other simulation modes? The design/implementation of new simulation modes should not impact the QoR (neither the runtime nor the memory consumption) of the existing nominal (serial) simulation (section V.C and VII)

e- Ease of use and simple validation flow (get rid of the complex DSF flow, see section III and V)

III. USF GENERAL CONCEPT

The Unified STIL Flow concept is based on the idea of using a unique STIL file to provide all possible simulation modes (serial, parallel, mixed serial/parallel and parallel with NShifts). The STIL file contains serial test patterns and it is the same file targeted for ATE testing.

From the validation flow point of view, the Unified STIL Flow allows for significant simplification in comparison to the old flow. Figure 2 depicts the different steps involved in both flows. The USF is simpler and straightforward, which allows for a robust and easy to use flow. The USF compliant STIL file (automatically generated by ATPG and does not require user interventions) is augmented with meaningful and compact information regarding the inserted DFT structures (see section IV). From an algorithmic point of view, the corresponding TB continues to derive the force/strobe test data from the original STIL for serial simulation (dark-blue path in fig 3). In addition, it performs a serial to parallel transformation of the load data and a parallel to serial transformation of the unload data for the parallel simulation (light-blue path in fig 3), as well as other appropriate test data transformations for the mixed serial/parallel and NShifts simulations modes (transformation process described in section V).

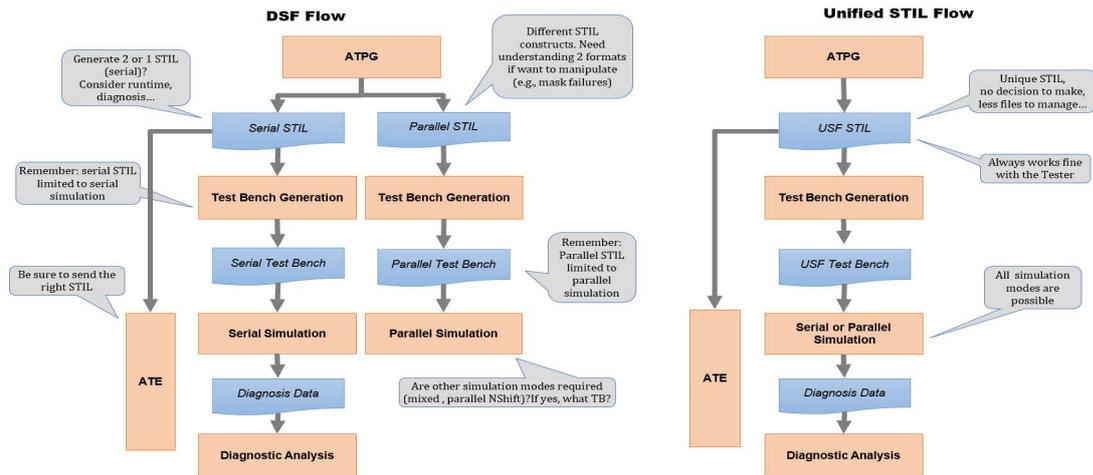


Fig 2: DSF Vs. USF

Therefore, unlike the DSF, the starting point is not a parallel DB, but the original (serial) DB at the primary SI/SO ports. Thus, the idea is to use the data-path in its normal direction (from serial to parallel in the decompressor, and from parallel unload to serial in the compressor) to avoid computing inverse CODEC functions (the compounding AND/OR/NAND gates, being symmetric functions, they are not reversible). This allows to derive parallel force load-data from serial, while in the DSF flow, we can simply not derive serial load from parallel data (opposite direction of the data-flow) as required by the NShifts simulation mode for instance. Another important aspect of the USF approach is the usage of high level of abstraction to model the DFT IP CODECs (namely the behavioral model) in the corresponding TB. They ensure serial/parallel load and parallel/serial unload transformations and serves for various purposes as well. First, they are used as a mean for validation of the DFT IP by providing another view. This in turn improves validation confidence since we are coding the same functionality using a different model between the netlist and the TB. Second, they allow for higher simulation performance (higher level of abstraction, faster than the netlist and RT levels). And finally, they allow for cycle compression for parallel sim. From a QoR point of view, we achieve a significant better performance through the concordance and collaboration of three different techniques: a) the usage of high level

model for DFT IP (behavioral model), b) the algorithmic optimization of the broad side technical for parallel force and strobe on all scan cells at the same time and, c) the cycle compression of sequential CODEC (from m to 0 or n cycles, where $n \ll m$). These advantages were confirmed by various real design experimentations (see section VII). Note that USF also addresses the limitations described in [8, s.5] for parallel and compressor/decompressor scan.

IV. STIL MODELING FOR USF FLOW

In the sequel, we describe the STIL based modeling required by USF approach in order to capture the description of the DFT IP. We assume that the user is familiar with such language [5]. A valuable effort has been provided to “unify” the interface for scan test generation based on STIL, by extending the usage of the STIL for various inputs [9], [13]. The result is a methodology adopted worldwide through an automated ATPG tool. On our side, we also extend the STIL, for pattern validation purposes. From one hand, we exploit ATPG spent time and computation effort to overcome deriving key knowledge of the DUT, and from the other hand, we simplify the flow by using one single input file (DB).

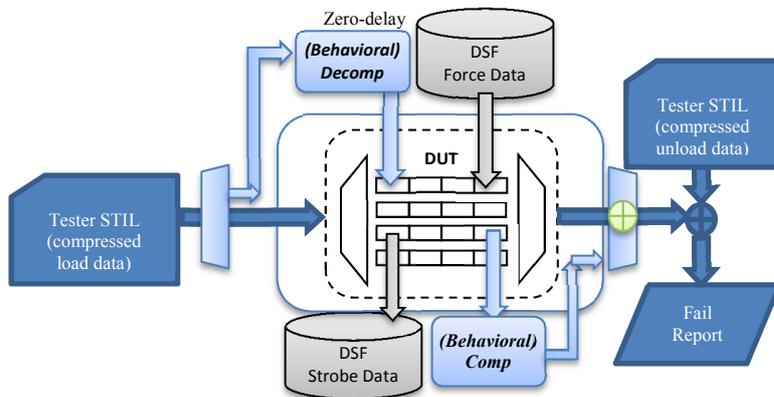


Fig 3: Data/Control flow diagram

Initially, the STIL was designed to solve the “gigabyte” problem, that is, to provide a way to transport large volumes of data efficiently, and at the same time, provide a standard that all ATE manufacturers can use. Therefore, at first glance, it may seem not a suitable format for our purpose. However, when careful studying the STIL language we observe that is composed of different sets of information. It holds test procedures and macros (e.g., `load_unload` and `capture` procedures), test data, timing information and test flow execution (*PatternBurst*). STIL has also a category of test structure that can be exploited to derive valuable information regarding the DFT solution, such as *ScanChains* and *BistStructure*. Unfortunately, this information is sufficient for certain test solutions (legacy scan, and LogicBIST) but falls short when used for a custom DFT IP. We need a mechanism to incorporate such missed information. To be efficient and coherent we want to avoid to vehicle it through yet another format and reuse at its maximum extent, the STIL file.

This task is not trivial, and requires first, detailed inventory of the additional needed information (mainly information related to the inserted DFT structure) and then, an efficient modeling of this information in the STIL file while respecting the overall STIL syntax. However, one of the main weakness of the STIL standard is precisely the hardware description (it was not meant for that). Even though, the STIL disposes of two essential data types to incorporate additional information as part of the test program: *Ann* (annotation) and *UserKeywords* (*UserFunctions*). Note that STIL doesn’t specify what to do with such data, up to the user to leverage its usage. Such flexibility was exploited to extend the STIL capability and model different DFT IPs. Hopefully, such extension was made easy by the need to describe only functional/behavioral information (we are concerned with the DFTIP CODEC behavior, not their exact hardware architecture which allowed for very compact models). On the other hand, we only need relevant data-flow part to be described. That is, only modules affecting the parallelization of the scan data such as the *Shift* procedure. Thus, only DFT structure that affect/modify the scan data need to be described. This is the case for the De-compressor, the Compressor and the X-Tolerant structure for instance. The modeling of complex OCC modules and JTAG-1500 interfaces can be avoided since during parallel simulation we rely on their nominal mode (drive their primary inputs) to bring their desired state for control flow.

By coupling such STIL extension with the usage of some CTL (Core Test Language) statements [10], such as *ScanMasterClock*, we were able to model in compact and compliant manner all required information.

The main blocks (all STIL standard compliant) required for DUT structure and the DFT IP behavior descriptions are:

- The *Header* block contains some generic but mandatory information like the DUT module name, the STIL format and the minimum parallel NShifts value.

- The *ScanStructure* STIL block groups all the scan chains, with the name, the scan input, the scan output and the scan cell data (cell name and polarity, !).
- The *CompressorStructures* block describes the Compressor and De-compressor DFT IP modules for the compressed designs. It specifies the load/unload pipeline stages and all the input and output connections with their dependencies with load, unload and mode input signals.

Figure 4 gives an example using these constructs.

```
Header { ...
  Ann { * top_module_name = top * }
  Ann { * Unified STIL Flow * }
  Ann { * min_n_shift = 0 * }
} ...
ScanStructures {
  ScanChain "1" {
    ScanLength 4; ScanInversion 1;
    ScanIn "top.S1.U1.Z" ScanOut "top.S1.U2.Q"
    ScanCells "top.S1.I13.T1" "top.S1.I14.T1" ! "top.S1.I23.T1" "top.S1.I24.T1" ;
    ScanMasterClock "CLK1" ; } ...
}
ScanChainGroups {
  "core_group" { "1"; "2"; ... "20"; } "load_group" { "test_si1"; "test_si2"; }
  "unload_group" { "test_so1"; "test_so2"; "test_so3"; "test_so4"; }
  "mode_group" { "test_si3"; } "enable_group" { "test_si4"; }
}
UserKeywords CompressorStructures;
CompressorStructures {
  Compressor "top_U_decompressor_ScanCompression_mode" {
    LoadPipelineStages 2;
    UnloadPipelineStages 3;
    ModeGroup "mode_group"; LoadGroup "load_group";
    CoreGroup "core_group";
    Modes 2;
    Mode 0 { ModeControls { "test_si3"=0; }
      Connection 0 ! "1" ! "4" ! ... "14" ! "15" ! "17" ! "20";
      Connection 1 ! "2" ! "3" ! ... "13" ! "16" ! "18" ! "19";
    }
    Mode 1 { ModeControls { "test_si3"=1; }
      Connection 0 ! "1" ! "3" ! "5" ! "7" ! "9" ! ... "15" ! "17" ! "19";
      Connection 1 ! "2" ! "4" ! "6" ! "8" ! "10" ! ... "16" ! "18" ! "20";
    }
  }
  Compressor "top_U_compressor_ScanCompression_mode" {
    UnloadGroup "unload_group"; CoreGroup "core_group";
    UnloadModeEnable "enable_group";
    Modes 9;
    Mode 0 { ModeControls { "test_si4"=0; }
      Connection "1" 0 1; Connection "2" 2; ... Connection "20" 2 3; } ...
    Mode 8 {
      ModeControls { test_si4=1; test_si3=1; test_si1=1; test_si2=1; }
      Connection "1" 0; Connection "2" 2; ... Connection "20" 3;
    }
  }
}
```

Fig 4: STIL model for Codecs

For instance, DFT solutions with sequential compression are described with the *SeqCompressorStructures* blocks that specify all the DFT IP sequential elements like the PRPGs, MISR...etc. In the presence of multiple Codecs this block is specified one per core.

V. SIMULATION MODES

Thanks to the USF approach, different simulation modes are enabled as explained below.

A. Serial (Nominal) Simulation

Serial simulation is the nominal simulation that mimics exactly the ATE operations according to the STIL data and protocol. It relies on true DFT IP implementation and the primary SI/SO interface (dark blue in figure 3). The USF approach of this mode reuse the same

implementation and behavior described in [1]. In addition, we need to make sure that the additional data-structure and algorithm for the new simulation modes are not negatively impacting the runtime performance of the nominal serial mode (memory impact cannot be avoided since we want to have all these modes at runtime to prevent long recompilation times, thus conditional compilation of certain code parts cannot be exploited). Such implementations aspects are too low-level details and cannot be described here, but a faithful indicator that they are correctly and efficiently conceived, is the runtime performance comparison of USF serial mode with the serial-only TB (section VII).

B. Parallel Simulation

To accelerate the simulation, the parallel scan mode applies the test vector in parallel in one single cycle rather than serially shifting in and out in n cycles. In the USF context, parallel simulation uses different path (light blue in fig 3) and relies on behavioral codec models to force and strobe scan cell data. The behavioral compressor/decompressor are 0 cycle optimized descriptions of their equivalent gate-level compressor/decompressor. The TB may add some extra cycles for NShifts or MBC (Multi-Bit-Cell) purposes (see section V.C).

The general parallelization procedure is as follows:

Step 1: identify opportunities for parallel accelerations protocol

Step 2: retrieve (from STIL) DFT CODEC functional and scan-chain structural descriptions

Step 3: derive accelerated behavioral models (FL and FU functions)

Step 4: generate a TB with both serial and parallel simulation capabilities (runtime simulation mode selection)

In the STIL file, the parallel acceleration is performed for the *Shift*, *Loop* and *LoopData* statements, where scan data is in play. So, in step 1, related *load_unload** Procedure holding these statements and manipulating scan-in and scan-out ports can be automatically detected and processed. Step 2 uses the modeling described in section IV to automatically extract the information of scan chains structure and internal scan cells SI and SO from the STIL *ScanStructure* block, to allow for parallel access of these internal nodes (the Verilog TB uses the *force/assign* statements to set the desired values). Likewise, the CODECs functions are retrieved from the STIL *ScanCompression* block. To our knowledge, there is no general-purpose deterministic method to derive behavioral models from GL or even RTL models. However, the derivation of accelerated models (step 3) can be fully automated if we reuse RTL modules developed by the DFT IP team. Nevertheless, our preferred approach is to use hand-written semi-automatic behavioral models. This allows to achieve better acceleration (better runtime simulation) than RTL

models from one hand, and a better validation confidence of the DFTIP by using a completely different implementation, from the other hand. This process is described hereafter.

Transforming the gate-level netlist behavior, cycle by cycle, to its equivalent optimized behavioral model that preforms the same logical/functional tasks is not a straightforward task, especially if we need acceleration by requiring far less clock cycles. Three categories exist. Those that require one single clock to advance the computation, thereby reducing n cycle to 1. Those that are pure computational and can be reduced to no cycle at all. And finally, those that are deeply sequential and requires m cycles where $m < n$ (total cycle required by the original GL netlist).

As seen before, the USF approach derives on the fly (do not rely on pre-computed ATPG DB) the parallel data from their serial counterparts. That is, it performs serial to parallel transformation of the load data to prepare for parallel force of all scan cells at once, and the parallel to serial transformation to the unload data to prepare for comparison with the golden data at the primary scan-outs. This simulation requires the knowledge of the transfer functions and the number, size and configuration of the scan chains. That is, we parse and interpret the STIL blocks describing decompressor (compressor) or any other transformation function between the primary scan-ins (scan-outs) and the actual scan chains. Figure 5 illustrates the basic idea. The (nominal) serial scan mode applies test data and strobes responses from the primary scan-ins and scan-outs (positions 0 and 3 in the figure). These test data are available also in serial format for the parallel scan mode. To derive their parallel data counterpart, USF proceeds as follows. The serial load data from $si[i]$ (position 0) are simulated using the decompressor function F1 to derive the internal values at position 1. Thereafter, the inversions between scan cells are considered and final parallel values at position "P" are derived. These values are applied in parallel (in one cycle) to the scan-ins of all scan-cells. In other words (assuming that all scan chains have the same length p), for each scan chain, the parallel load data $IP_0 IP_1 \dots IP_{p-1}$ are derived from the serial load data $IO_0 IO_1 \dots IO_{p-1}$ using the function FL:

$$(IP_0 IP_1 \dots IP_{p-1}) = FL(IO_0 IO_1 \dots IO_{p-1}) \quad (1)$$

$$FL = F1 \circ M' \quad (2)$$

M' is input inversion mask vector: $M'_i = (\sum_{k=0}^{i-1} mk) \% 2$

where the FL is the composition function of the decompression function F1 with the input inversions mask function. The elements of vector M' , $mk = 1$, if cell- k has an inverted output, 0 otherwise. Note that hierarchy levels (scenarios of DFT CODEC insertion at top level then routed to block level) not shown in figure 5, have neutral polarity in general. In case of inverted polarity (odd number of inversions) such information can easily be ported in the inversion masks of cell 0 / cell $n-1$ respectively or within the Decompressor/Compressor descriptions.

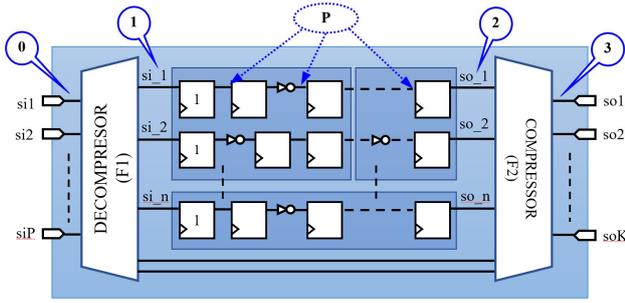


Fig 5: Compressed Scan Block Diagram

To derive the parallel load test data, the FL function can be implemented using a software approach. That is, the FL transformation is applied to the serial patterns to derive the parallel patterns internally to the TB Generator. These patterns are then generated in the *testb.dat* file. This file would contain both serial and parallel load/unload data. This approach allows a simple test protocol file (*testb.v*) that can be easily shared for both serial and parallel scan modes. It presents however several limitations: redundant information (load/unload data represented twice although in two different formats), bigger test data file and ambiguous STIL \leftrightarrow Verilog TB correspondence since the STIL file is now storing only serial information. To address these limitations, the adopted approach relies on a unique test data file (holding the serial data only), interchangeably used for serial and parallel scan simulations. The FL function is implemented at HDL level using behavioral descriptions, in the Verilog test bench file (the *testb.v*). When the logic simulation starts, the testbench reads in the (serial) test data from the test data file and transforms these data into parallel ones prior to their application on the scan cells. The HDL implementation of the FL function doesn't consume additional simulation cycles, so that the testbench timing is not perturbed. This approach is also used for the Compressor as explained in the sequel. The F1 (de-compression function) is provided in the STIL file under *CompressorStructures* block as shown in (figure 4).

The expected (unload) data are only available in serial format (at position 3 in figure 5). To allow comparing the response data against these expected data, a parallel-to-serial transformation of the response data is required. This transformation is naturally made by the DFT structure itself but would take p cycles to finish. In parallel mode, we want to accelerate the simulation, so it takes one clock cycle at most (no cycle for combinatorial compressor) for the "unload" as well. Let $q_1(p)$, $q_2(p)$.. $q_n(p)$, the outputs of scan cells nearest to SOs in chains 1, 2 ..n respectively. Likewise, the outputs of the second stages (second scan cells from the SO) will be $q_1(p-1)$, $q_2(p-1)$.. $q_n(p-1)$, and so on. The unload data (at position 3) are given as function of data at positions "P" in figure 5. That is:

$$\begin{aligned} u(p) &= FU(q_1(p), q_2(p).. q_n(p)) \\ u(p-1) &= FU(q_1(p-1), q_2(p-1).. q_n(p-1)) \\ &\dots \\ u(1) &= FU(q_1(1), q_2(1).. q_n(1)) \end{aligned}$$

$$\begin{aligned} \text{where: } FU &= M'' \circ F2 \quad (3) \\ M'' \text{ is output inversion mask vector: } M''i &= (\sum_{k=i}^{n-1} mk) \% 2 \end{aligned}$$

That is, we first apply the inversion function M'' to the outputs of the scan cells to retrieve their value at position 2 (internal scan-outs) and apply thereafter, the F2 (compression) function to retrieve the values at position 3 (primary scan-outs). Then, for each scan output, these DUT responses $u(p)$ $u(p-1)$.. $u(1)$ are compared against the expected (serial) data given in the STIL file.

The concept is the same for a combinatorial compression (w/wo XTOL), or sequential compression (using MISR blocks for instance), only F2 function needs to be described accordingly. As for the de-compressor description, we retrieve the descriptions of the compressor module from *CompressorStructures* block in the STIL file (see fig 4, section IV). For instance, in case of XOR (spatial) compression, the testbench file applies a bit-level XOR function simultaneously to all scan cells outputs, so that in the same cycle, data seen at the outputs of scan cells (position "P") are transformed (serialized) to data at primary scan-outs (position 3) in 0 simulation cycle (behavioral model).

When other transformation functions are present, they are also considered to derive the unload function. For the specific case of XTOL scan compression,

$$FU = M'' \circ F\text{-XTOL} \circ F2 \quad (4)$$

where F-XTOL function masks the possible X value of certain internal scan-outs, prior to their compression.

Note finally that for sake of efficient TDV management, we do not perform the strobe data process on scan-cell .so rather on .si nodes by following the permutation:

$$so(i) = si(i+1) XOR m(i) \quad (5)$$

where $m(i) = 1$ when there is an inversion between cell i and cell $i+1$, 0 otherwise, provided in the STIL *ScanStructure* block. Thus, we hold only .si nodes (instead of .si and .so nodes) which allows to greatly reduce the size (storage and manipulation) of STIL files and the TB as well. This in turn has positive impact on the simulation runtime. Such manipulation is always performed except for few specific cases where we need to have separate si and so representations (not treated in this paper).

C. Partial Parallel

A key benefit of the USF approach is its ability to provide an efficient solution for one of the critical problems in pattern validation, namely NShifts support and mixed serial/parallel for scan compression designs. Under the DSF flow, these modes are not possible using only the parallel single STIL.

The process of accelerated validation through a parallel simulation requires a direct access to the scan cells in order to overcome the time-consuming serial scan shifting. When non-scan sequential elements exist in the design (e.g., shadow latches), they are not modeled in the STIL/CTL since they are not part of the DFT structure. Therefore, the pattern validation testbench doesn't know how to get direct access to these resources. As a result,

these sequential elements remain initialized and affect the general behavior (expected values) of the DUT, completely corrupting the validation process. To cope with this limitation, the solution consists to get access to these non-scan elements through their normal data path, that is, serially through the outputs of the neighbor scan FFs. This comes to combine the parallel simulation with some serial shift cycles. The overall simulation is still mainly done in parallel, so that the time acceleration benefit is preserved, but in this case all sequential elements of the DUT can be initialized correctly and thus a reliable pattern validation can be ensured. We refer to this capability with “NShifts”, meaning that for a given parallel scan simulation, there will be $NShifts$ bits (equals to the length of deepest shadow logic) serially shifted-in for each test pattern. Figure 6 depicts its basic principle. Consider a scan chain with p cells and at max, ns shadow cells. The parallel load data bit at position i are b_i for all i in $[0..p-1]$. The NShifts mechanism is composed of two steps. The parallel data b_i are first (computationally) shifted left by ns (value of NShifts) so that they are parallel applied to cell positions “ $i-ns$ ” for i in $[ns..p-1]$. The second step consists to perform “ ns ” serial shift, so that load data b_i for i $[0..ns-1]$ are stored in the first ns cells, while in the same time all the previously loaded (in step 1) are serially shifted by ns positions as well. At the completion of the second step, all b_i values are stored in their exact cells locations i for all i in $[0..p-1]$. Figure 6 gives an example for such procedure where $ns=2$, $p=5$ and the first and second load steps are highlighted in blue and green respectively. Note that for each (parallel) value b_i we need to compute the possible ns backward inversions (due to ns shifts) so that at the completion of the final step the right b_i polarity is stored at position i .

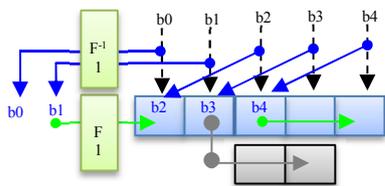


Fig 6: Principle of the NShifts simulation

One of the main limitations of such an approach is encountered when there is a transfer function like the De-Compressor module in case of compressed scan (module F1 in figure 5). In the second step, the derivation of the serial data bits $b_0 b_1 \dots b_{ns-1}$ from their parallel counterpart, requires the simulation of the inverse function $F1^{-1}$. Depending on the complexity of the base function F1, this function can be very hard to compute or simply impossible (NAND and NOR gates are symmetric functions and thus do not have inversion functions). To cope with this important limitation, the NShifts procedure is performed differently thanks to the USF flow. First, we start from the serial data (rather than the parallel data). The first step consists to derive the parallel $p-ns$ load data (by considering the possible ns backward inversions), that is, the bits $b_{ns} b_{ns+1} \dots b_{p-1}$, and apply them to the scan cells $[0..p-1-ns]$. The remaining $b_0 b_1 \dots b_{ns-1}$

load data do not need any computation (serially ready). The second step consists then to just serially shift in these data in the scan chains (to be loaded at scan cells $[0..ns-1]$). In other words, in the new approach, the NShifts procedure can be seen as partially parallel scan simulation applied to $p-ns$ bits, followed by the conventional serial (using the serial ready load data) mode simulation for the remaining bits. It starts from the serial data to easily derive the partial parallel data. That is, we simulate the De-Compressor function in the right serial-to-parallel direction, on the contrary of the previous approach that starts from the parallel data and needs to simulate the inversion function of the De-Compressor. An immediate consequence of the adopted approach is that, not only the NShifts feature can be provided for compressed scan, but also suffice to reuse the same (serial) STIL file for different NShifts values (runtime dynamic change to avoid time-consuming TB recompilations).

Mixed Serial/Parallel sim: The USF flow also allows to greatly simplify the mixed serial/parallel simulation mode. This capability is important to bypass time consuming *test_setup* procedure (can reach several hours per test session), and gathering in the same simulation session, the report for all simulated patterns. In a mixed serial/parallel simulation scheme, we specify the set of patterns that need to be serially simulated; the remaining are parallel simulated. Since USF starts from the same (serial) STIL file, the serial simulation doesn't require any additional computation, whereas the parallel simulation activates the transfer functions mentioned previously. That is, mixed serial/parallel simulation is starting from the same (serial) STIL file and consists to simply enabling/disabling the FL and FU transformation functions depending on the actual setting of the present pattern while taking care of various FSM states to prepare the required data one patterns before the pivoting pattern. Thus, the USF approach allows treating the scan structures from their primary I/Os so that only serial information is required whatever the requested simulation technique (serial, parallel, parallel with NShifts, or mixed serial/parallel).

VI. USF CHALLENGES AND SOLUTIONS

Besides the main goal of functional test patterns validation, it's desirable that the USF concept efficiently tackle other validation aspects.

The first challenge concerns timing simulation. The primary concern of parallel simulation is the zero-delay simulation in order to “functionally” validate the test patterns. It ensures similar validation as the ATPG good simulation machine, while ensuring higher confidence since it's handled by an ATPG independent system (the Logic Simulator). The timing order and event placements are depicted in figure 7. The default parallel operation will place the force event 1 time increment before the first defined clock pulse in the shift vector. This is the latest possible time which allows the device scan state to stabilize on the scan elements. Likewise, it places release

events at end of the cycle to allow for enough hold time. While this schedule accommodates all zero-delay simulation and some timing simulation scenarios as well, it cannot always prevent the violation SDF timing constraints. For instance, when setup violations have occurred, the simulation may generate X values on the flops and affect the simulation response for the next unload operation. Similarly, for situations where the clock distribution results in large internal delays relative to the input timing, we need to delay the release event sometimes even beyond the current cycle to capture the right force value. To account for such circumstances, the TB holds special Verilog parameters (*par_force_time* and *par_release_time*) to specify a different parallel force and release event placements. They can be both positive or negative to respectively delay or advance their positions with respect to their nominal placements. In all cases, parallel timing simulation requires a carefully thought to avoid the risk of violations. An event too close to the clock edge, and a timing checks may be violated with X mismatch on that scan element. Conversely, an event set too earlier may also cause mismatches if the scan state has not stabilized.

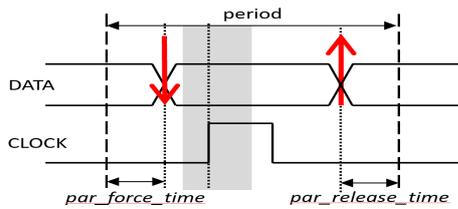


Fig 7: Parallel Simulation Timing Control

The proposed parallel scan simulation approach can be seen as parallel load/unload-serial compare technique. In comparison to the parallel load/unload – parallel compare technique used by the previous DSF approach. The new technique is more precise in the sense that it is closer to the actual serial procedure performed by the tester. Also, it tests the real scan-out data expected at the primary SO while considering possible Compressor failures masking. However, unlike the old (full) parallel approach, it doesn't allow precise identification of the failing scan cell/chain from compressed scan architectures. Nevertheless, it allows for ATPG tool diagnosis by providing a good datalogger of the failing patterns similar to the serial sim. Besides this capability, we propose two USF based ATPG-free debug strategies allowing various debug resolution/runtime and memory tradeoffs. The advanced debug mode relies on the modeling of additional DFTIP structure in the STIL file with possible comparison between ATPG and simulation values of these internal nodes (equivalence checking principle). The intermediate debug mode uses an ATPG generated optimized binary file (only strobe data coded on 2 machine bits rather than 1 Byte ASCII char, thus 1/8th of original DSF scan data size) holding data of selected observation points before the compressor (improve observability principle). For sake of space, these two modes are targeted for a future paper.

VII. IMPLEMENTATION AND EXPERIMENTAL RESULTS

The USF approach was implemented under an automation framework [1], [11]. Several tens of parameters allow various TB configurations. An excerpt from example of the generated Verilog TB, showing some behavioral parts for a Sequential Compression solution is given below.

```
// MAX TB Test Protocol File
// Module under test: des_unit
// Generated from original STIL file : "patterns.stil"
// STIL file version: "1.0"
// Simulation mode: default parallel simulation
...
`define CHAINOUT0 {dut.S1.I1.LOCKUP.Q, ... }...
`define CHAININ0 {dut.S1.I1.I4.TI, ... }...
...
module USF_testbench ();
...
design dut (.out ( in_con ), out ( out_com ), ...
    .clk ( clk_con ), .test_se ( test_se_con ), ...
    .test_si1 ( test_si1_con ), ..., .test_so1 ( test_so1_con ), ... );
...
task reseed_with_misr_load_unload ; ...
task shadow_to_careprg ; ...
task shadow_to_xtolprg ; ...
task load_unload ;
begin ...
    if (parallel_sim) begin
        generate_prp(); generate_xprp(); generate_signature;
        p_Loop_0(idargs, valargs);
    end else Loop_0(cnt_ScanCompression_mode);
end endtask
// serial to parallel pattern data translation
task generate_prp; begin
    while (pidx < prpcount) begin
        PRPQ = {PRPD[0], PRPD[98:1]} ^ (PRPD[0] ? 99'b0000010101000...000:
99'd0); ...
        LODCHP0[psci] = PRPS[84] ^ PRPS[47] ^ PRPS[24]; ...
    end end endtask
task generate_xprp; begin ...
    while (pidx < xprpcount) begin ...
        XPRPQ = {XPRPD[0], XPRPD[98:1]} ^ (XPRPD[0] ? 99'b0000010...000: 99'd0);
        XTOLCTRL = XPRPQ[87] ^ XPRPQ[39] ^ XPRPQ[22];
        if (XTOLCTRL) begin SXTOLOUT0[shpsci] = SXTOLOUT0[idx]; ...
    end else begin
        SXTOLOUT0[shpsci] = XPRPQ[91] ^ XPRPQ[60] ^ XPRPQ[9]; ... end
    if (xnew_seed != 1) begin
        XTOLOUT0[xpsci] = SXTOLOUT0[idx]; ...
        xpsci = xpsci + 1; end end endtask
// parallel to serial data translation using the behavioral model
MaxTB_SEQCOMPRESSOR comp0 (comp_en, comp_in, comp_sel, comp_sel_b,
xtolen_w, comp_out);
task generate_signature; ... begin
    UNLODCHP0 = `CHAINOUT0 ^ {CH_OUTINV[0], 1'b0}; ...
    for (pidx=mpsci; pidx < sigcount; pidx=pidx+1) begin
        comp_in = {UNLODCHP33[pidx], UNLODCHP32[pidx], ...
UNLODCHP0[pidx]};
        comp_sel = {XTOLOUT8[pidx], XTOLOUT7[pidx], ... XTOLOUT0[pidx]};
        xtolen_w = XTOLEN[pidx];
        feed_and_state = {1'b0, misrq[31:1]} ^ ({32{misrq[0]}} &
32'b100001000101010...001);
        next_state = feed_and_state ^ comp_out; misrq = next_state;
    end end endtask endmodule
// snippet of the behavioral compressor module
module MaxTB_SEQCOMPRESSOR (comp_en, chin, sel, sel2, xtolen, dout);
task compute_xmask; begin
    if (mode==1) xmask = {34{1'b0}}; else xmask = {34{1'b1}};
    casex (xsel) 9'b011010111, 9'b00???????, ... 9'b1111?????: xmask[0] = mode;
endcase ... end endtask
    assign xsel = {sel[0], sel[1], sel[2], sel[3], sel[4], sel[5], sel[6], sel[7], sel[8]};
    always @ (posedge compen) begin
        if (xtolen == 1'b0) begin chout = chin;
        end else begin
            compute_xmask (xsel, 1, xmask); chout = xmask & chin;
        end end
    assign dout[0] = chout[6] ^ chout[8] ^ chout[26]; ...
endmodule
```

Fig 8: Some TB Coding Blocks

The table below presents an example of USF parallel optimizations per STIL statement, for one test pattern.

STIL test patterns	Sim Vectors	
	Serial	USF
<i>pattern 2:Call reseed_with_misr_load_unload {si4=0111001010001; ... so1=LLLLLHLL; ...}</i>	13 V	0 V
<i>Call shadow_to_careprpg;</i> <i>Call reseed_partial_overlap_load_unload { cnt:=0; si4=0111001010100;...; }</i>	1 V 13 V	0 V
<i>Call shadow_to_xtolprpg;</i> <i>Call reseed_partial_overlap_load_unload { cnt:=2; si4=0011110100111;... }</i>	1 V 13 V	0 V
<i>Call shadow_to_careprpg;</i> <i>Call load_unload { cnt:=200; }</i>	1 V 200 V	1 V
<i>Call multiclock_capture { pi=P.....; po=LLL.....; }</i>	1 V	1 V
<i>Total /pattern</i>	243 V	2 V

TABLE I: SIMULATION CYCLES COMPARISON

We observe that for USF parallel simulation many STIL statements do not consume any simulation cycles leading to a cycle compression ratio > 100x.

On the other hand, table 2 reports the experimentation results for 1000 patterns in USF, DSF and nominal serial simulation modes. It concerns a Scan Compression design with 1 million scan cells with maximum scan chains length of 256. In comparison to DSF, USF allows for a 2X runtime improvement and 100x disk space saving.

	STIL size	TB size	TB data	Comp (secs)	Sim (secs)
Serial	124M	4M	20M	1447	194200
DSF	2.2G	200M	2G	1474	4382
USF	124 M	194M	20M	1447	2124

TABLE II: PERFORMANCE EVALUATION

Hence, we observe a net improvement of USF over DSF for both runtime and memory. Besides, by validating the real protocol executed by the Tester, USF ensures better validation confidence. Compared to serial simulation, USF parallel simulation is 1 or 2 degrees of magnitude better, and in this particular testcase the ratio is 91X. The ratio between serial and USF parallel simulation will increase with the maximum scan chain length. The table 3 (below) summarizes the main advantages and disadvantages for each simulation technique.

Property	Serial	DSF	USF
ATPG runtime	=	--	=
ATPG memory	=	=	=
Simulation runtime	---	--	+
Simulation memory	+	--	=
Disk allocation	+	--	+
Validation confidence	+++	-	+
Parallel with NShifts	+	NA	+
Mixed serial/parallel	NA	NA	+
Ease of USE	=	=	+
TTM	=	=	-
Diagnosis	+	NA	+

TABLE III: SIMULATION FLOWS COMPARISON

VIII. CONCLUSION

We presented a non-intrusive (no DFT IP modification is required, and no test patterns and data are altered) validation technique called Unified STIL Flow. It's an effective methodology to accelerate simulation and significantly reduce validation time. At the contrary of the previous DSF approach, the parallel acceleration did not incur coverage loss of DFT IP (i.e., while testing the DUT, the generated ATPG patterns test also the inserted DFT IP compression logic). It also allowed for a simple and efficient mechanism to ensure parallel validation with NShifts as well as mixed serial/parallel simulation for DFT design using either combinational or sequential scan compression. The USF methodology presents good flexibility that allows for runtime switching between these simulation modes to overcoming time-consuming recompilation phases. Another important benefit of the USF is its ease of use by providing a simple and robust flow with as reduced entry points as possible (single STIL file for validation, debug and diagnosis). In comparison to existing flows, it presents better QoR simpler flow and better coverage.

REFERENCES

- [1] S. Boutobza, S. Popa, A. Costa " An Automatic Testbench Generator for Test Patterns Validation ", EWDTS 2018.
- [2] S. Boutobza, S. Popa, A. Costa "A Journey from STIL to Verilog ", paper 182, EWDTS 2018
- [3] U. Schoettmer, T. Minami "Challenging the "high performance-high cost" paradigm in test", ITC 1995
- [4] W.R. Simpson "Cutting the cost of test; the value-added way", ITC 1995
- [5] "Standard Test Interface Language (STIL) for Digital Test Vectors", IEEE 1450.99
- [6] B.J. Oomman, W.T Cheng J. Waicukauski "A universal technique for Accelerating Simulation of Scan Test Patterns", pp 135-141, ITC 1996
- [7] Synopsys "STILDVP User guide" version N-2017.09, March 2018
- [8] R. Raghuraman, "Simulation requirements for vectors in ATE formats", pp. 1100-1107, ITC, 2004
- [9] "Elements of STIL: principles and applications of IEEE Std. 1450" par Gregory A. Maston, Tony R. Taylor, Julie N. Villar, Springer Ed.
- [10] "CTL for Information of Digital ICs", par Rohit Kapur, Kluwer Academic Publishers
- [11] Synopsys "TetraMAX Test Pattern Validation User guide" version P-2019.03, March 2019
- [12] <http://en.wikipedia.org/wiki/ATPG>
- [13] P. Wohl ; J. Waicukauski, "A unified interface for scan test generation based on STIL ", pp. 1011-1019, ITC, 1997

An Accelerator-based Architecture Utilizing an Efficient Memory Link for Modern Computational Requirements

Saba Yousefzadeh¹, Katayoon Basharkhah¹, Nooshin Nosrati¹, Rezgar Sadeghi¹, Jaan Raik², Maksim Jenihhin², Zainalabedin Navabi¹

¹School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

²Department of Computer Engineering, Tallinn University of Technology, Estonia

{saba.yousefzadeh, basharkhah.kt96, nosrati.nooshin, rr.sadeghi, navabi}@ut.ac.ir, {jaan.raik, maksim.jenihhin}@ttu.ee

Abstract— Hardware implementation of many of today’s applications such as those in automotive, telecommunication, bio, and security, require heavy repeated computations, and concurrency in the execution of these computations. These requirements are not easily satisfied by existing embedded systems. This paper proposes an embedded system architecture that is enhanced by an array of accelerators, and a bussing system that enables concurrency in operation of accelerators. This architecture is statically configurable to configure it for performing a specific application. The embedded system architecture and architecture of the configurable accelerators are discussed in this paper. A case study examines an automotive application running on our proposed system.

Keywords— Heterogeneous Systems, Accelerator-based Architecture, Hardware Accelerator, On-Chip Communication Architectures.

I. INTRODUCTION

Existing embedded systems provide a handful of embedded processors, a convenient bussing structure, and a memory system. Embedded systems can be configured and/or programmed for implementation of applications with algorithms that are generally procedural, and require some computations. Some embedded environments allow custom instructions to be defined to handle complex and heavy computations in hardware instead of using processor instructions. Custom instructions, elaborate bus structures, and near-processor memory in some embedded system can be used to improve the performance of executing certain applications. However, they fail to provide efficient execution of repeated calculation of arithmetic operations and the concurrency in the execution of these calculations.

Kernels that form the heart of automotive, bio, security, telecommunication and many of today’s applications for IoT and machine learning consist of repeated looping of multiply, add, divide, and some basic arithmetic operations. These kernels perform functions like FIR filtering, trigonometric functions, Fast Fourier Transform (FFT), Cordic, and many other such functions.

Such applications, not only require efficient execution of various kernels that they consist of, but they also require concurrent execution of these kernels. As mentioned, existing embedded systems have limited success in the efficient execution of the aforementioned applications.

For improving performance of the execution of such applications, efficient execution of kernels, and concurrency in their execution must be addressed. These two requirements

are the focal point of the embedded system that we are discussing in this paper and will be elaborated further in the following paragraphs.

Efficient execution of kernels requires hardware accelerators that, not only have a good performance for a set of like kernels but also can be adapted to other kernels with similar arithmetic requirements. This means that ideally, we want to have several accelerators that can be configured to accelerate a class of related kernels. This configuration requirement calls for an accelerator engine with a datapath and a controller, both of which can be programmed or configured. The datapath is to be configured for the utilization of datapath components, such as multipliers, adders, and dividers. On the other hand, the controller is configured for the flow and sequencing in which these components are utilized. This paper proposes a configurable accelerator for a certain class of kernels.

The other requirement that must be considered is concurrency in the execution of accelerators running different kernels. To achieve this, an embedded system consisting of several accelerators with an efficient mechanism for data exchange between concurrently running accelerators must be devised. In addition, the processor of such an embedded system must be able to exchange data with the individual accelerators.

Other features of this embedded system are a bussing system that can be used for processor-memory data transfer, as well as processor-accelerator data exchange.

For a reconfigurable system, the embedded system must a) facilitate configuration of the individual accelerators, and b) configure the flow of activation and data transfer of the accelerators. In such a system the main embedded processor can be responsible for configuring the flow, or a sequencer attached to the main processor can take this responsibility.

Although the focus of this paper is on configurable accelerators. The embedded environment in which these accelerators are embedded is also of our concern. We will present the overall architecture of our embedded system. We show how the embedded procedural processor works with the accelerators and the memory structure through our proposed bussing and communication structures.

In this paper, an accelerator-based embedded architecture is proposed. The accelerators take advantage of computation configurability, and configurable communications between accelerators and between accelerator and memory. The

interconnect infrastructure is based on a combination of fast and slow busses and a DMA for inter-accelerator data transfer.

The section that follows, presents some of the previous works on accelerator rich architectures, while in Section III, an overall view of the accelerator-based architecture is depicted. Section IV shows the dedicated accelerator’s design and it’s configuration. Section V explains the details of the configurable communication system. A case study will be studied in Section VI. The last section presents our conclusions.

II. Previous Works

Many research works are done on accelerator-based architectures in the recent decade. Work has also been done on the required architectures for different applications such as machine learning and neural networks [1], graph analytics [2], smart vision [3] and bioinformatics [4] that consists of kernels such as FFT, Convolution and Sobel filter. Also, there are many heavy DSP and classification processes in IoT sensor nodes that require energy efficiency.

Most recent works, focus on CGRAs as an alternative accelerator, providing more efficiency in comparison to FPGAs by adding reconfigurability to accelerator design. Reference [5] proposes a reconfigurable Integrated Programmable Array (IPA) accelerator added to the PULP heterogeneous cluster that computational kernels would map on the IPA. In such designs, processing elements would be configured to do different required kernels in the application. There are many architectures including dedicated hardware accelerators for a specific application like bio-signal analysis [6], [7]. In these designs, CPU and accelerators execute different segments of an application. In a way, CPU some time acts as a sequencer or task scheduler.

Another view is the flexibility of the Accelerator rich architectures that is the concern of many works. CHARM is a heterogeneous architecture including accelerator building block and an accelerator block composer (ABC) that dynamically connects different ABBs together for different functionalities. This gives tremendous flexibility in composing different accelerators at different times, thereby better-matching application demand [8].

III. Accelerator-Based Processing System

This section presents the architecture, configuration, and operation of our proposed accelerator-based embedded system. The architecture is designed to handle computationally heavy applications in today’s technology. Applications such as ADAS, machine learning and bio-system applications that require computations involving large repetition of array-multipliers, dividers, adders, and subtractors will particularly benefit from our proposed architecture. In such computations, repeated instances of arithmetic units are replaced with iterations that repeat the use of a limited number of instances of the arithmetic units. A processing unit handling such computations is referred to as an accelerator. An accelerator executes a kernel of an application.

The system described here is a configurable accelerator-based embedded system. It is configurable in the sense that

computations of the individual processing elements and communications between them must be configured for a specific application before it is ready to actually execute the application. It is accelerator-based because many configurable accelerators handle the hard jobs of heavy computations. The system is embedded, because it contains a processor that handles the main task of execution of a task by running the sequential parts and scheduling and assigning iterative parts concurrently to the accelerators.

In addition to running the applications, the processor is responsible to configure the system before running the application begins. This system configuration is static, and it is fixed for the entire duration of running an application. In what follows, we will present the top-level architecture of our system, followed by a configuration procedure, and then the operation of this system when running an application.

A. Top-Level System Architecture

As with any processor-based computing engine, our system has the three distinct parts of processing elements, memory, and processor-memory communication. These parts are described in the subsections that follow. Figure 1 shows the top-level structure of our embedded system.

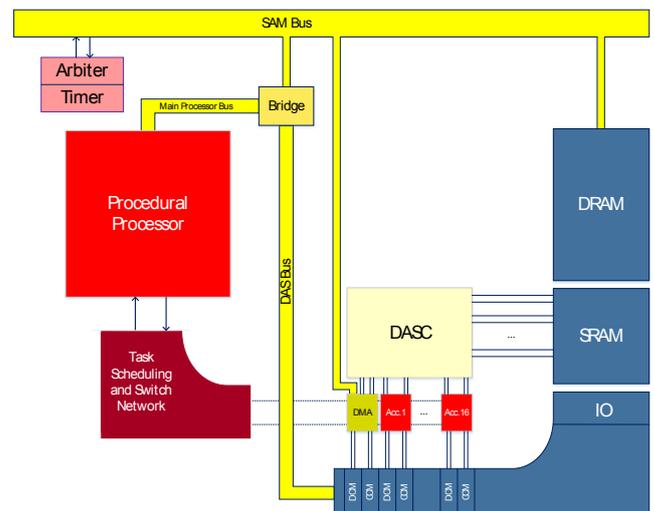


Figure 1. The top-level structure of the proposed embedded system

1) Processing Elements

Shown in light red are the processing elements of our system. Shown in darker red is a configurable switching network that is configured to handle task scheduling among the processing elements (light red). There are two categories of processing elements. One is the main processor that a) configures the system, b) executes sequential parts of an application, and c) handles data exchange between the processing elements. The processor component is a standard processor with instructions for data movement, jump, and branch, basic logic, and arithmetic, memory and IO device operations. The other category of processing elements in this system are the accelerators. These engines have a limited set of instructions and generally handle arithmetic operations in nested loops. The accelerators are configurable for the type of instructions they execute, and the arithmetic units and internal accelerator busses that they utilize. Once configured

before an application begins, they remain unchanged throughout the execution of the application.

2) Memory System

The other part of any computing system is its memory. Our proposed embedded system has two types of memory, a slow DRAM structure, and a synchronous SRAM. This is shown in Figure 1 in blue. Both memory types will use the same addressing and are mapped in different address segments. The system being memory-mapped, IO devices and accelerator configuration memories also use the same addressing with different mappings.

The DRAM occupies most of the address space of the embedded system. This part of the memory is connected to an arbitrated bus that is used by the main processor and a DMA. The DMA is part of a bussing system that is responsible for the transfer of data between the DRAM and SRAM.

The SRAM part of the memory is connected to a data router and switching network that connects it to the accelerators and to the DMA for processor access and data transfer to and from DRAM. This is a fast memory and operates with the same clock as the processor. This SRAM is primarily for data that is used by the accelerators.

The memory map of our system has another SRAM that is used for DMA data and accelerator instruction and configuration. This is a dual-port memory accessed through the processor address space as well as the DMA or the accelerators. This part of the memory is shown by the horizontal extension of the memory in Figure 1.

3) The Bussing System

Shown in yellow in Figure 1 is the communication system of our embedded system. This part consists of an NoC switch box, two busses, and a DMA for communication between two busses that, in effect, facilitates data transfer between the SRAM and DRAM. The two busses use the same address bus and connect via a bridge.

The upper bus in Figure 1 is called SAM Bus that stands for Slow Arbitrated Memory Bus. As shown, this bus extends vertically to the DMA. As requested by the processor, the DMA connects to DRAM through this bus for transfer of data to SRAM.

The other bus in this system is DAS Bus that stands for Synchronous Accelerator Direct Bus. This is a fast bus and is synchronized with the processor clock. The only master of this bus is the processor and unlike the SAM Bus, no arbiter or timer is needed for it.

B. Configuration Procedure

The architecture shown in Figure 1 facilitates configuration of the accelerators and the other parts of the system before the execution of an application begins. In this pre-execution configuration, the main processor reads configuration information from DRAM and using the bus bridge connects SAM Bus to DAS Bus for the configuration information to be written into the corresponding SRAM segments. One of the components that are being configured is the task scheduling and switch network (dark red) component that is considered as part of the processing elements of our embedded system.

Once the configuration of components is complete, control returns to the processor to begin execution of the application for which the embedded system is configured.

IV. Accelerator Architecture

Figure 2, shows a cluster of accelerators and some details of each accelerator. In general, accelerators can be implementing a complete application, or they can be used collectively to form a complete application. Our point of view of an accelerator is the latter. These accelerators can implement a simple task like multiply and accumulate to more complex one like Fast Fourier Transform, Filters, Trigonometric Functions, Cordic, and Convolution. These tasks become kernels of, for instance, an automotive application, e.g., vehicle detection and tracking, and together with the procedural processor and the bussing structure of Figure 1, they implement the planned application.

When designing such accelerators different challenges arises. Although parallel computing provides good flexibility, the overhead of fetching instructions [9], controller-datapath overheads and also multiple load-store in the processor-memory links decreases the energy efficiency and performance of such systems. The accelerator architecture can remove a majority of these issues by revisiting register transfer level architecture and configuration of the accelerators. In the next subsections, these issues are discussed.

A. RTL Architecture

Like any RTL design, the accelerators contain a datapath and a controller. The datapath of a typical accelerator in the proposed architecture is shown in Figure 3(a). There are arithmetic and data buffering units inside the datapath. The arithmetic unit consists of add and multiply units that can be utilized for the formation of many iterative MAC operations. Some multiplexers are embedded in the architecture to decide on the inputs of the arithmetic unit. The source of this data can be either neighboring accelerators or the main memory. Data buffering unit are input/output channels for memory interfacing and communications. These buffer channels can be a FIFO/ random access buffer, register file or even dual-port register file.

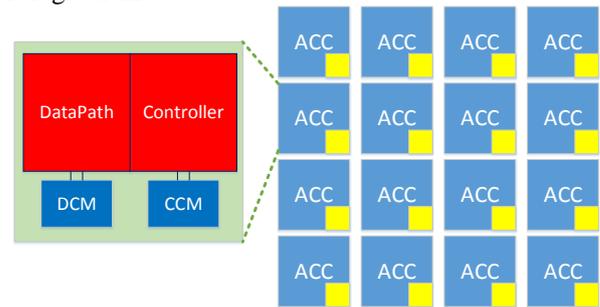


Figure 2. Accelerator configurable architecture

For the control unit, a state machine like the partial one shown in Figure 3(b) enables the corresponding control signals based on the application.

B. Configurable Architecture

Configuration in an accelerator-based architecture can be static or dynamic. In dynamic mode, accelerators perform different tasks being reconfigured in each computation cycle. However, this method adds an overload of fetching multiple instructions and increases power consumption. Our solution to this problem is a local configuration or instruction memory. These configuration memories can be filled before

the process execution. As shown in Figure 2 there are two configuration memories, one for data and the other for control signals. We only perform static configuration in this work.

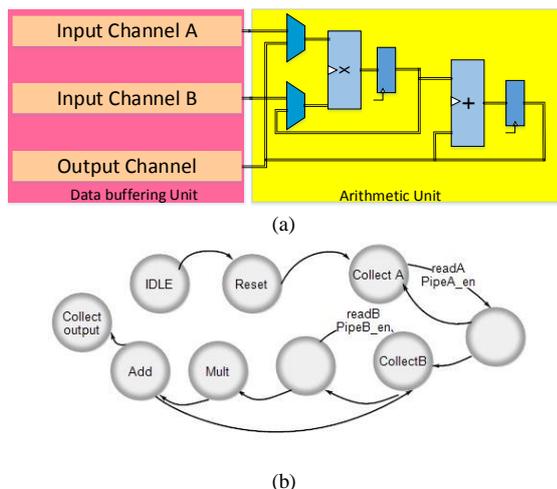


Figure 3. RTL architecture (a) datapath (b) controller

There are different kinds of data in an application. Constant data, variable data that must be read and written consecutively and the input-output arrays that must be accessed externally. Application constant data like the coefficient of a filter or the number of iterations for internal loops are stored inside DCM (Data ConFIGuration Memory). In the proposed architecture, there is a need for indexing the input-output buffers to access data and for enabling different registers inside datapath. These indexing values are also stored in DCM. Data configuration memory is a dual-port memory-mapped SRAM. Data locality gets the computing unit and the data storage closer to each other and helps in achieving the high-efficiency challenge in an accelerator-based system.

Another dual-port memory-mapped SRAM is dedicated as the local configuration for control signals. This is called Control Configuration Memory, CCM. Instructions and the instruction sequencing are stored statically inside this memory. The controller configuration can be provided by programming a micro-sequencer to adopt the operation of the machine to work as different kernels of an application like as a filter or trigonometric processor.

V. Communication Infrastructure

They are several de-facto standards for a communication architecture in embedded systems including the Advanced Microcontroller Bus Architecture (AMBA), Avalon, CoreConnect, STBus and Wishbone [10]. Each of these busses is defined based on different topologies, arbitration methods, bus-width, and types of data transfer to handle the connection between components of an embedded system in a specific manner. However, they cannot fit into our accelerator-based embedded system. Therefore, we design our communication structure with a generic style based on which that for each application can be configured.

This section presents the communication infrastructure of our accelerator-based embedded system, shown in yellow in Figure 1. Our communication architecture consists of three distinct parts: (A) Bussing structure including three busses

coupled by a bridge, (B) Direct Memory Access (DMA), and (C) Dedicated Arbitrated SRAM communication (DASC). Following subsections illustrate these parts.

A. Bussing Structure

The bussing structure is organized into two bus segments, SAM Bus (Slow Arbitrated-Memory) and DAS Bus (Directed-Accelerator Synchronous). These busses are connected to the main bus of the procedural processor via a bridge.

The SAM bus is a low-speed, multiple-master, central arbitrated bus which is intended to connect the procedural processor and DMA to DRAM. The processor (through the bridge) and the DMA unit are the only bus masters. When they request control of the SAM bus, the arbiter module examines each request to grant the highest priority using arbitration mechanisms specified by the configured bus protocol. The SAM bus does not operate with the processor clock and a timer module handles the bus timing specifications. It consists of address, data, and control lines and also supports burst and pipelined transfers.

The DAS bus is a high-speed, single-master, synchronous bus that allows connection to the procedural processor, and to configuration memories of DMA and the accelerators. The configuration memories are considered as memory-mapped on this bus, which is no different than the I/O devices. The bus has a memory-mapped interface that is only driven by the processor through the bridge module and supports synchronous data transfer. Because of its single master, this bus does not need an arbitration module. Like the SAM bus, DAS consists of address, data, and control lines and also supports burst and pipelined transfer.

The bridge acts as a slave device on the main processor bus and a master for SAM and DAS bus. It is initiated by the procedural processor to perform address decoding and then establish interconnection of the main processor bus to SAM or DAS bus based on decoded address. The bridge module also provides buffering of all address, data and control signals to handle different data rates.

Table 1 summarizes the comparison among the busses of our bussing structure.

Table 1 – Comparison among busses

SAM	DAS	Main processor bus
Low-speed	High-speed	High-speed
multiple-master	single-master	single-master
asynchronous	synchronous	synchronous

B. DMA

The Direct Memory Access (DMA) module collaborates in both SAM and DAS busses to allow access to the memory system independent of the procedural processor. It becomes a bus master on SAM to coordinate memory traffic between DRAM and the procedural processor. The DMA module, as a slave on the DAS bus, serves the processor requests to gain access to SRAM. It is also responsible for transferring data from SRAM to DRAM. For this data transfer, at first, the procedural processor initiates the DMA module through the

Table 2. Comparing execution time

	Susan_Smoothing	Susan_Edge	Susan_Corner	Total
Sequential	143 ms	15 ms	10 ms	168 ms
Proposed Architecture	57 ms	4 ms	2.8 ms	62 ms

VII. Conclusions

In this paper, a configurable accelerator-based embedded system has been proposed to improve the performance of traditional processor-based systems. The architecture is designed to handle computationally heavy applications by assigning iterative tasks concurrently to accelerators.

Our architecture consists of a procedural processor and several configurable accelerators that are connected to each other and a memory system through a communication infrastructure. The processor executes procedural tasks of a specific application while the nested loops, regarded as kernels of an application, are performed by the accelerators. The accelerators are configured for the computation they perform.

The communication infrastructure includes a bussing structure, DMA and a routing and switching network. This provides a high-speed, high-performance interconnect architecture for integration of our embedded system's components.

Before starting an application, both the processing elements and communication between them must be configured for that specific application.

The MiBench benchmark is run on the accelerator-based embedded system to evaluate the performance of the system where the accelerators are involved. Experimental results show a 2.7 X speedup.

REFERENCES

- [1] Y. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits*, 52(1), 127-138, 2016.
- [2] T. J. Ham, L. Wu, N. Sundaram, N. Satish, M. Martonosi, "Graphiconado: A High-Performance and Energy-Efficient Accelerator for Graph Analytics," In Proc. of 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pp. 1-13. IEEE, 2016.
- [3] S. Das, D. Rossi, K. J. M. Martin, P. Coussy, and L. Benini, "A 142MOPS/mW Integrated Programmable Array accelerator for Smart Visual Processing," In Proc. of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-4, 2017.
- [4] S. Huang, et al. "Hardware Acceleration of the Pair-HMM Algorithm for DNA Variant Calling," In Proc. of the ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 275-284, 2017.
- [5] S. Das, K. J. M. Martin, P. Coussy, and D. Rossit, "A Heterogeneous Cluster with Reconfigurable Accelerator for Energy Efficient Near-Sensor Data Analytics," In Proc. of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5, 2018.
- [6] A. Roy, et al. "A 6.45 Self-Powered SoC With Integrated Energy-Harvesting Power Management and ULP Asymmetric Radios for Portable Biomedical Systems," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 9, pp. 862-874, 2015.
- [7] X. Liu, et al. "An Ultralow-Voltage Sensor Node Processor With Diverse Hardware Acceleration and Cognitive Sampling for Intelligent Sensing," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, pp. 1149-1153, 2015.
- [8] J. Cong, et al. "Accelerator-Rich Architectures: Opportunities and Progresses," In Proc. of the 51st Annual Design Automation Conference (DAC), pp. 1-6, 2014.

- [9] M. Gautschi, P. D. Schiavone, A. Traber, I. Loi, A. Pullini, D. Rossi, E. Flamand, F. K.Gürkaynak, and L. Benini. "Near-threshold risc-v core with dsp extensions for scalable iot endpoint devices. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*," PP(99):1-14, 2017.
- [10] M.Mitić, and M.Stojčev, "An overview of on-chip buses", *Facta universitatis-series: Electronics and Energetics*, Vol.19, No.3, PP.405-428, 2006.
- [11] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, R. B. Brown, "MiBench: A free, commercially representative embedded benchmark suite," In Proc. of the Fourth Annual IEEE International Workshop on Workload Characterization, (pp. 3-14), 2001.

Antenna Array Calibration Algorithm Based on Phase Perturbation

Victor Djigan
Institute for Design Problems in
Microelectronics of RAS
National Research University of
Electronic Technology
Moscow City, Russian Federation
djigan@yandex.ru

Vladislav Kurganov
Institute of Microdevices and
Control Systems
National Research University of
Electronic Technology
Moscow City, Russian Federation
kurganov@org.miet.ru

Abstract—The paper presents an algorithm of antenna array calibration, which estimates and compensates phase lags, caused by non-identical electrical characteristics of the array channels. To estimate the phase lags, the algorithm uses only the array output power measurements under the specific channel phase perturbations. The estimation accuracy equals phase shifter quantization step and does not depend on the number of the array channels. The algorithm is compared with two similar calibration algorithms, known from publications, whose accuracy depends on the number of the array channels and is much less compared with that of the proposed algorithm. For this reason, the algorithm, presented in this paper, can be widely used for efficient arrays calibration, signal source angular position estimation and tracking by a calibrated or a non-calibrated array with any aperture shape: linear, flat or conformal and with an arbitrary channel, selected as a reference one.

Keywords—array calibration, signal source tracking, phase perturbation, power measurements

I. INTRODUCTION

Today a lot of modern radio systems use antenna arrays as directional antennas, because such antennas provide non-mechanical beam steering, output Signal-to-Noise Ratio (SNR) improvement, interference signals suppression, multi-beam operation and some others. In millimeter wave applications, due to small geometrical sizes and progress in semiconductor and micromechanical technologies, it became possible to integrate the receivers, transmitters and antennas of the arrays into inexpensive devices, called “antenna-on-chip” and “antenna-in-package”, which are manufactured as inseparable electronic components [1]. Unfortunately, such arrays often require calibration, which cannot be provided manually like in arrays with discrete components, because there is no mechanical access inside the integrated devices. In this case, the antenna-on-chip and antenna-in-package calibration can be provided only by using the external control of the required parameters. Calibration means the adjustment of the array channels characteristics making the channels “electrically identical” each other to ensure the far-field coherent addition of the received/transmitted microwave signals according to the fundamental principle of the antenna arrays operation. The calibration is achieved by the channel gains and phase lags equalization.

To equalize channel gains, an array has to contain the digitally controlled attenuators or amplifiers. Such devices are not used extensively, because the attenuators decrease received/transmitted power and the microwave amplifiers with variable gain are complicated in design. However, if channel gain variation is just a few dB, this does not affect significantly the shape of the array radiation pattern [2]. In this case, the gain equalization is not such a critical option. That is why the gain equalization is not considered in this paper.

The situation is quite different in the case of the transmission lines, which connect the array antennas with its receivers and transmitters. Because the mechanical design constraints, it is rarely possible to ensure the same length of the lines, especially in the integrated antenna-in-chip and antenna-on-package devices. Besides, because the variation of electrical characteristics of the used materials, the phase lags can be different even in the same channels of the different samples of the array. The length variation is the reason of the additional random $-\pi \dots \pi$ phase lags in array channels. Phase lags variation leads to significant distortion of array radiation pattern shape right down to a complete destroying [3]. Channel phases can be also changed, if semiconductor amplifiers of the antenna arrays change characteristics due the external radiation. Thus the radiation resistant amplifier design and array channel phase lag equalization are mandatory ones.

Calibration is usually provided for each manufactured array by the preliminary estimation of the channel phase lags, the lag values storing and further using in the calculation of phases, set by phase shifters for beam steering. If the antenna array is used to track a signal source location, which is determined by the spatial phase lags from the source to the antennas of the array, then the estimated spatial and channel lags can be compensated simultaneously, providing the array calibration and beam steering towards the signal source. This might be done by means of the same phase shifters, which are inherent devices of any antenna array.

This paper considers a new array calibration/signal source angular position estimation algorithm. For each channel, the algorithm uses only the array output power measurements under the specific phase perturbations in only pairs of the antenna array channels: reference and calibrated ones. The target of the paper is to provide the algorithm description and

to evaluate the estimated channel phase lags accuracy and Euclidian distance between the radiation pattern of the calibrated array and that of the array, which does not require calibration. This accuracy is compared with that of similar calibration algorithms [4, 5].

II. CALIBRATION ALGORITHM OVERVIEW

The simplest phase calibration algorithms are based on the search of the optimal phase values in antenna array channels, which maximize the array received/transmitted power. Such algorithms are very similar to the phase adaptation algorithms [6]. However, it is known, that the cost functions in the phase adaptation tasks are multiextremal [7] that does not guarantee a unique solution of a calibration task, even by using the discrete phase shifters with a small phase quantization step [8]. Advanced calibration algorithms use phase measurements, which require the expensive measuring instruments. Such calibration is basically provided in laboratory conditions.

A survey of a number of advanced antenna array calibration algorithms is presented in [9]. Among known calibration algorithms, the most attractive are the ones, which do not require phase measurements, access to channel signals or channel disconnection [10]. These algorithms use a limited number of the only antenna array output power measurements, which allow to extract (estimate) the channel phase lag values in un-calibrated arrays, spatial phase lags in calibrated arrays or both ones simultaneously. This allows to compensate the estimated phase lags, ensuring the coherent addition of the channel signals. Such calibration is inexpensive and allows to provide the calibration even during the array operation. The examples of such algorithms are [4, 5]. Unfortunately, these algorithms are not so accurate. That is why a new calibration algorithm has been developed.

III. PROPOSED CALIBRATION ALGORITHM

This paper presents an array calibration algorithm, which is similar to algorithm [4], but requires a simpler procedure of a channel phase perturbation and provides a better accuracy. The paper demonstrates simulation results, which confirm the statement.

The proposed calibration algorithm is developed using the following steps. The signal at the m -th channel output of the un-calibrated array is described as

$$u_m(t) = A(t) \exp(j\omega t) k_m \exp[j(\psi_m + \delta\psi_m + \varphi_m + \delta\varphi_m)], \quad (1)$$

where $A(t)$ is the modulation (information) signal; ω is the angular frequency of the carrier; t is the time; k_m is the real-valued channel gain; ψ_m is the spatial phase lag, caused by signal source orientation for the receiving array or a receiver orientation for the transmitting array; $\delta\psi_m = -\pi \dots \pi$ is the random channel phase lag; φ_m is the current value of channel phase, caused by a phase shifter;

$$\delta\varphi_m = -\pi/2^B \dots \pi/2^B \quad (2)$$

is the uniformly distributed phase shifter quantization error; B is the number of the bits of the digitally controlled phase shifters and M is the number of the array antennas.

In this case, using trigonometric relationships, it is easy to show, that the antenna array output power is described as

$$P = P_0 \sum_{m=1}^M \sum_{n=1}^M k_m k_n \times \cos(\psi_m + \delta\psi_m + \varphi_m + \delta\varphi_m - \psi_n - \delta\psi_n - \varphi_n - \delta\varphi_n), \quad (3)$$

where

$$P_0 = E\{|A(t)|^2\} \quad (4)$$

is the power of the signal, received/transmitted by one antenna of the array and $E\{\bullet\}$ is the averaging operator.

To extract the spatial ψ_m and channel $\delta\psi_m$ phase lags, it is required to conduct a number of the antenna array output power measurements under the specific channel phase perturbations, i.e. changing the channel phases by phase shifter as

$$\varphi_M^{(\text{new})} = \varphi_M^{(\text{cur.})} + \Delta\varphi_M, \quad (5)$$

where

$$\varphi_M^{(\text{cur.})} = [\varphi_1, \varphi_2, \dots, \varphi_{m-1}, \varphi_m, \varphi_{m+1}, \dots, \varphi_{M-1}, \varphi_M]^T \quad (6)$$

is the vector of the current (unperturbed) phase values, caused by phase shifters, and $\Delta\varphi_M$ is the perturbation vector, selected from the below ones:

$$\Delta\varphi_M^{(11)} = [0, 0, \dots, 0, 0, 0, \dots, 0, 0]^T_M, \quad (7)$$

$$\Delta\varphi_M^{(12)} = [\pi, 0, \dots, 0, 0, 0, \dots, 0, 0]^T_M, \quad (8)$$

$$\Delta\varphi_M^{(k1)} = [0, 0, \dots, 0, \Delta\varphi_m, 0, \dots, 0, 0]^T_M, \quad (9)$$

$$\Delta\varphi_M^{(k2)} = [\pi, 0, \dots, 0, \Delta\varphi_m, 0, \dots, 0, 0]^T_M. \quad (10)$$

Here, $\Delta\varphi_m|_{k=2} = \pi$, $\Delta\varphi_m|_{k=3} = -\pi/2$, $\Delta\varphi_m|_{k=4} = \pi/2$, the subscript M denotes the number of the elements in a vector and the superscript T denotes the vector transposition. So, the proposed phase perturbations (5) – (10) require the changing phase shifter values only in a pair of array channels: a reference (1-st) and a considered (m -th) per power measurement.

Then, using trigonometric relationships, it is possible to show that the following relationships

$$\begin{cases} p_m^{(11)} - p_m^{(12)} - p_m^{(21)} + p_m^{(22)} = 8p_0 k_1 k_m \cos(\tilde{\psi}_m) \\ p_m^{(31)} - p_m^{(32)} - p_m^{(41)} + p_m^{(42)} = 8p_0 k_1 k_m \sin(\tilde{\psi}_m) \end{cases} \quad (11)$$

are valid, where

$$\tilde{\psi}_m = \psi_m + \delta\psi_m + \varphi_m + \delta\varphi_m - \psi_1 - \delta\psi_1 - \varphi_1 - \delta\varphi_1 \quad (12)$$

is the estimated total phase lag in the m -th channel of the non-calibrated array, which includes the fixed phase lag $\psi_1 + \delta\psi_1 + \varphi_1 \neq 0$ of the reference channel and the unknown phase shifters errors $\delta\varphi_1$ and $\delta\varphi_m$ of the reference and m -th channels.

Thus, according to (11), using eight measurements of the antenna array output power under channel phase perturbations (5) – (10), i.e. in pairs channels only, the total phase lag in the m -th channel relatively the antenna input of the reference channel can be extracted from the measurements as

$$\tilde{\psi}_m = \arctan \frac{p_m^{(31)} - p_m^{(32)} - p_m^{(41)} + p_m^{(42)}}{p_m^{(11)} - p_m^{(12)} - p_m^{(21)} + p_m^{(22)}} + k\pi, \quad (13)$$

where $k = 0, \pm 1$ and $k\pi$ is a correction term to the arctangent, taking into account the location on the complex plane of a complex value, the real part of which is the denominator and the imaginary part is the numerator of the presented fraction.

The procedure (5) – (10), (13) has to be executed for all channels of the array, except a reference one. The measuring of $p_m^{(11)}$ and $p_m^{(12)}$ powers can be done only once and then used in (13) for each $m = 2, \dots, M$ channel. Thus, the proposed phase estimation algorithm (5) – (10), (13) requires $6(M-1) + 2$ measurements of the array output power.

The equation (13) is the same as that of algorithm [4], where another channel phase perturbation procedure is used. In the referred algorithm, antenna array output power $p_m^{(11)}$ is measured under phase perturbation

$$\Delta\varphi_M^{(11)} = [0, 0, \dots, 0, 0, 0, \dots, 0, 0]_M^T, \quad (14)$$

power $p_m^{(12)}$ is measured under phase perturbation

$$\Delta\varphi_M^{(12)} = [0, \pi, \dots, \pi, \pi, \pi, \dots, \pi, \pi]_M^T \quad (15)$$

and powers $p_m^{(k1)}$ and $p_m^{(k2)}$ are measured under phase perturbations

$$\Delta\varphi_M^{(k1)} = [0, 0, \dots, 0, \Delta\varphi_m, 0, \dots, 0, 0]_M^T \quad (16)$$

and

$$\Delta\varphi_M^{(k2)} = [0, \pi, \dots, \pi, \Delta\varphi_m, \pi, \dots, \pi, \pi]_M^T, \quad (17)$$

where values $\Delta\varphi_m|_{k=2} = \pi$, $\Delta\varphi_m|_{k=3} = -\pi/2$ and $\Delta\varphi_m|_{k=4} = \pi/2$ are used in the vectors $\Delta\varphi_M^{(k1)}$ and values $\Delta\varphi_m|_{k=2} = 0$, $\Delta\varphi_m|_{k=3} = \pi/2$ and $\Delta\varphi_m|_{k=4} = -\pi/2$ are used in the vectors $\Delta\varphi_M^{(k2)}$.

Comparing to (5) – (10), phase perturbations (5), (6), (14) – (17) are more complicated, because they require the phase changing in all the array channels accordingly (15) and (17). Besides, in [4], the phase shifter quantization errors $\delta\varphi_1$ and $\delta\varphi_m$ are not taken into consideration. So, with the perturbations vectors (14) – (17) the equation (13) is valid only if the phase shifter quantization errors are absent in all the channels, except the reference and m -th ones, that cannot be guaranteed for all channel simultaneously and leads to significant errors in phase lags $\tilde{\psi}_m$ estimation comparing with the algorithm, considered in the presented paper. The algorithm [4] also requires $6(M-1) + 2$ measurements of the array output power.

A similar channel phase estimation algorithm

$$\tilde{\psi}_m \approx \arctan \frac{p_m^{(31)} - p_m^{(41)}}{p_m^{(11)} - p_m^{(21)}} + k\pi \quad (18)$$

is presented in [5], where the amendment $k\pi$ to arctangent is the same as in (13). This algorithm requires only $3(M-1) + 1$ measurements of the array output power under (5) – (7), (9) phase perturbations in pairs of the array channels.

However, (18) shows that the calibration algorithm [5, see p. 520] is only approximately estimates $\tilde{\psi}_m$ values. Because of this, the estimation accuracy of the algorithm is lower even that of [4] and that of the algorithm, considered in the presented paper. This will be demonstrated in the following section.

To calibrate array or (and) to steer the array beam, $\tilde{\psi}_m$ values have to be compensated by phase shifters as

$$\varphi_m^{(\text{new})} = -\tilde{\psi}_m + \varphi_m, \quad (19)$$

where $\varphi_m^{(\text{new})}$ are the new values of channel phases, set by phase shifters. For the calibration, it is possible to assume that $\psi_1 + \delta\psi_1 + \varphi_1 = 0$ for the reference channel, because the same additional phase lag $\psi_1 + \delta\psi_1 + \varphi_1 \neq 0$ in all channels does not affect the radiation pattern shape.

If accuracy of (13) or (18) is acceptable, any of the three considered algorithms with an arbitrary selected reference channel can be used for the calibration of the arrays with any aperture shape, because the phase lags $\tilde{\psi}_m$ are estimated for each array channel.

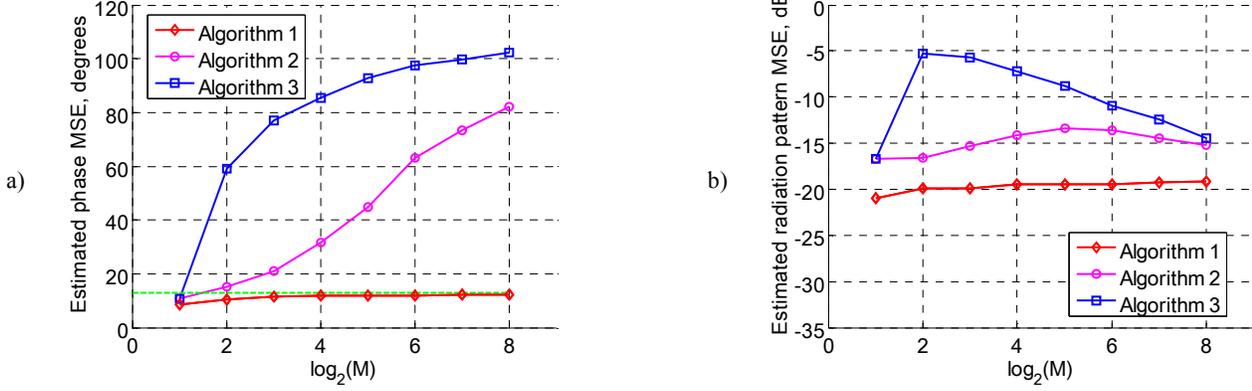


Fig. 1. Calibration accuracy $B = 3$: a) phase estimation MSE (green line is $\text{MSE}^{(\text{Theor.})} = 13^\circ$); b) radiation pattern MSE

IV. SIMULATION RESULTS

It follows from (12), that an error of channel lag estimation by the considered in this paper algorithm, is determined as

$$\delta\tilde{\psi}_m = \delta\phi_m - \delta\phi_1. \quad (20)$$

Because $\delta\phi = -\pi/2^B \dots \pi/2^B$, from (20) it follows that the proposed algorithm error of channel phase lag estimation varies in the range

$$\delta\tilde{\psi}_m = -\pi/2^{B-1} \dots \pi/2^{B-1}. \quad (21)$$

So, the maximal error $|\delta\tilde{\psi}_{\max}|$ of the calibration algorithm equals the phase shifter quantization step $\pi/2^{B-1}$, i.e. it is twice bigger than the maximal phase shifter quantization error $|\delta\phi_{\max}| = \pi/2^B$. Because the errors $\delta\tilde{\psi}_m$ are uniformly distributed between $-\delta\tilde{\psi}_{\max}$ and $\delta\tilde{\psi}_{\max}$ values, the Mean Square Error (MSE) of the $\tilde{\psi}_m$ estimation is determined as

$$\text{MSE}^{(\tilde{\psi})} = \sqrt{(\delta\tilde{\psi}_{\max})^2/12}. \quad (22)$$

Fig. 1a) shows the MSE values for the digitally controlled discrete phase shifters with typical number of bits $B = 3$ for all the three considered algorithms. Here calibration Algorithm 1 is one proposed in this paper, Algorithm 2 is that of [4] and Algorithm 3 is that of [5]. The values in the figure plots are averaged over 1000 experiments, conducted with randomly generated variables $\delta\psi_m = -\pi \dots \pi$, $\delta\phi_m = -\pi/2^B \dots \pi/2^B$ and $k_m = -1 \dots 1$ dB for the linear antenna arrays with $M = 2, 4, 8, \dots, 256$ channels and the same distances $d = \lambda/2$ between adjacent antennas. In Fig. 1a), the dashed horizontal line marks the theoretical value of Algorithm 1 MSE, which accordingly (21) and (22) is

$$\text{MSE}^{(\text{Theor.})} = \sqrt{(\pi/2^{B-1})^2/12}. \quad (23)$$

Similarly to Fig. 1a), for the different number of phase shifter bits and independently the number of antenna array channels, the phase estimation accuracy of Algorithm 1 approaches to the following values of MSE: $B = 2$, $\text{MSE}^{(\text{Theor.})} = 26^\circ$; $B = 3$, $\text{MSE}^{(\text{Theor.})} = 13^\circ$; $B = 4$, $\text{MSE}^{(\text{Theor.})} = 6.5^\circ$ and $B = 5$, $\text{MSE}^{(\text{Theor.})} = 3.25^\circ$.

Fig. 1a) demonstrates, that Algorithm 1 accuracy depends on the number of phase shifters bits B only and does not depend on the number of channels M . At the same time, accuracy of Algorithm 2 depends on the both B and M values, because due to phase perturbations (5), (6), (14) – (17), the errors $\delta\phi_m$ are different in the same channels during different power measurements. As a result, the accuracy of $\tilde{\psi}_m$ estimation is decreased as M is increased, because (13) becomes an approximate. Algorithm 3 accuracy does not depend on B and is determined by the approximate form of (18) only. The accuracy is also decreased as M is increased.

The phase estimation accuracy can be also evaluated in terms of averaged normalized Euclidian distance $\text{MSE}^{(\tilde{F}(\theta))}$ between the radiation patterns of the array, whose new phase shifts are set as (19), and the array, which does not require calibration. The distance is defined as

$$10 \log_{10} E \left\{ \frac{\sum_{n=1}^N [\tilde{F}(\theta_n) - F(\theta_n)][\tilde{F}(\theta_n) - F(\theta_n)]^*}{\sum_{n=1}^N F(\theta_n)F^*(\theta_n)} \right\}, \quad (24)$$

where $\tilde{F}(\theta_n)$ is the radiation pattern value towards the θ_n direction of the array, which uses new values $\phi_m^{(\text{new})} = -\tilde{\psi}_m + \phi_m$, set by phase shifters after calibration; $F(\theta_n)$ is the radiation pattern value towards the θ_n direction of the antenna array, which has equal phase lags and uses

phase shifter values $\varphi_m = -\psi_m$. In (24), N is the number of the points of the radiation pattern calculation over a range of $\theta_{\min} \dots \theta_{\max}$ angles.

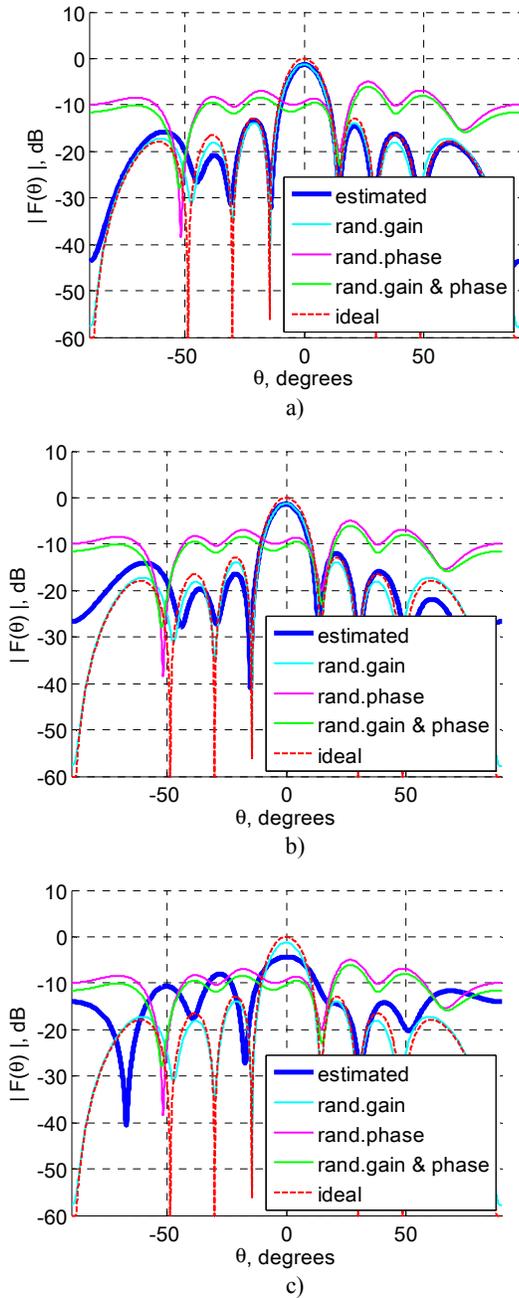


Fig. 2. Radiation pattern examples: a) Algorithm 1; b) Algorithm 2; c) Algorithm 3

The examples of radiation patterns from the above considered experiments for the linear arrays with $M = 8$, $B = 3$ and $\theta = 0^\circ$ are presented in Fig. 2. Here, the legend “estimated” refers to the radiation pattern of an array with estimated by the considered algorithms total channels phase lags, which are compensated as (19); the legend “rand.gain” refers to radiation pattern of an array with gain

errors only; the legend “rand.phase” refers to the radiation pattern of an array with channel phase errors only, the legend “rand.gain & phase” refers to the radiation pattern of an array with both channel phase errors and the legend “ideal” refers to the ideal radiation pattern of an array with no channel gain and phase errors.

The graphs in Fig. 2 pictures also confirm, that the dominated influence on the array radiation pattern shape is caused by the channel phase errors $\delta\psi_m$. The radiation patterns of the calibrated array are approached to those of arrays with channel gain errors only. The Euclidian distances between these particular radiation patterns are about -18 dB for Algorithm 1, -12 dB for Algorithm 2 and -6 dB for Algorithm 3 that is close to the averaged values in Fig. 1b). The calibrated array radiation pattern shapes are the confirmations of a better accuracy of Algorithm 1 compared with that of Algorithm 2 and Algorithm 3.

V. CONCLUSION

Therefore, the most accurate algorithm among the three considered ones is Algorithm 1. It provides an array calibration if the values $\psi_m = 0$ and $\varphi_m = 0$ are ensured during power measurements. This allows to estimate and store the pure channel phase lags $\delta\psi_m$ values, which has to be taken as the correction values during the array beam steering. If $\psi_m \neq 0$ and $\varphi_m \neq 0$, then the algorithm can be used to track the signal source location or calibrate and track simultaneously.

REFERENCES

- [1] Y.P. Zhang and D. Liu., “Antenna-on-chip and antenna-in-package solutions to highly integrated millimeter-wave devices for wireless communications,” IEEE Trans. Antennas and Propagation, vol. 57, pp. 2830–2841, October 2009.
- [2] A. Sebak, L. Shafai, H. Moheb, A. Ittipiboon, “The effect of random amplitude and phase errors on phased arrays performance”, Symposium on Antenna Technology and Applied Electromagnetics, pp. 391–396, 1990.
- [3] Ya. S. Shifrin, Statistical antenna theory, Golem Press, 1971.
- [4] M. K. Leavitt, “A phase adaptation algorithm,” IEEE Trans. Antennas and Propagation, vol. 24, pp. 754–756, September 1976.
- [5] R. Sorace, “Phased array calibration,” IEEE Trans. Antennas and Propagation, vol. 49, pp. 517–525, April 2001.
- [6] H. Steyskal, R. A. Shore and R. L. Haupt, “Methods for null control and their effects on the radiation pattern,” IEEE Trans. Antennas and Propagation., vol. 34, pp. 404–409, March 1986.
- [7] D. V. Nezlin and V. I. Djigan, “Structure of cost functions in discrete phase adaptation of antenna arrays,” Electronic Engineering. Series 10: Microelectronic Devices, vol. 75, pp. 3–6, No. 3, 1989. (in Russian)
- [8] V. I. Djigan and D. V. Nezlin, “Gradient algorithms in discrete phase adaptation of antenna arrays,” Radioengineering, pp. 86–87, No. 5, 1991. (in Russian)
- [9] E. V. Korotetsky, A. M. Shitikov and V. V. Denisenko, “Phased antenna array calibration methodths,” Radioengineering, pp. 95–104, May 2013. (in Russian)
- [10] G. G. Bubnov, S. M. Nikulin, Yu. N. Serebryakov and S. A. Fursov, Perturbation method of phased antenna array parameter measurements, Moscow: Radio and Communication, 1988. (in Russian)

Power Supply Noise Rejection Improvement Method in Modern VLSI Design

Melikyan Vazgen Sh.,
Synopsys Armenia Educational
Department
Yerevan, Armenia
vazgenm@synopsys.com

Mkhitarian Artur Kh.,
Synopsys Armenia Educational
Department
Yerevan, Armenia
marthur@synopsys.com

Kostanyan Hakob T.,
Synopsys Armenia Educational
Department
Yerevan, Armenia
hakobk@synopsys.com

Grigoryan Hayk T.,
Synopsys Armenia Educational
Department
Yerevan, Armenia
hgrigo@synopsys.com

Kostanyan Harutyun T.,
Synopsys Armenia Educational
Department
Yerevan, Armenia
harutyk@synopsys.com

Grigoryan Mushegh T.,
Synopsys Armenia Educational
Department
Yerevan, Armenia
mushegg@synopsys.com

Musayelyan Ruben H.,
Synopsys Armenia Educational
Department
Yerevan, Armenia
musayely@synopsys.com

Margaryan Hayk V.,
Synopsys Armenia Educational
Department
Yerevan, Armenia
hmargar@synopsys.com

Abstract— In modern integrated circuits the technology scaling and supply voltages values decreasing caused degradation of noise immunity of VLSI ICs. Therefore, the rejection of the noises in the power rails become a huge challenge considering the fact of area and power consumption requirements of the modern ICs. This paper presents a power supply noise rejection improvement method based on the combination of MOS (metal oxide semiconductor) and MOM (metal oxide metal) devices as capacitors which purpose is the decrease of dependency of decoupling capacitor capacitance on PVT (process, voltage, temperature) variations. According to the achieved results design of decoupling capacitors based on the proposed method decreases the total occupied area by 1,5 times without degradation of the noise immunity of the circuits.

Keywords— decoupling capacitor, noise immunity, metal oxide metal capacitor, metal oxide semiconductor capacitor, supply noise, IC design,

I. INTRODUCTION

In modern integrated circuits (IC) the channel length of transistors has reached up to 5nm. It gives opportunity to increase device density on die and provide more functionality in unit area. Also increased the noise influence on circuits performance. That's why the usage of high noise resistant designs are relevant. To overcome this problem various methods were used. The most common one is usage of decoupling capacitors, which opposes the quick voltage changes. They filter out the noise spikes and pass through only DC part of signal.

By placing decoupling capacitor between supply rails gives opportunity to decrease the noise level between them. Decoupling capacitances can be placed either out of IC or inside them. The decoupling placed outside of IC attenuates the noise between interconnections of I/O's and constant supply sources. And the decoupling placed inside the IC ensures high noise immunity of separate parts of circuit (Fig. 1) [1].

According to the (1)

$$|Xc| = 1/2\pi fC \quad (1)$$

Where:

Xc -capacitive resistance

f -signal frequency

C -capacitance

as bigger the capacitance value as lower the reactive resistance, hence the noise level between rails is also lower.

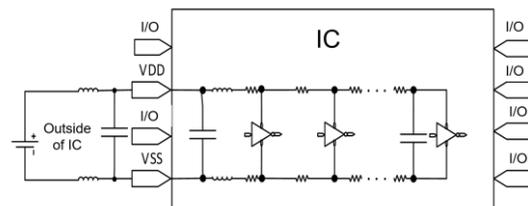


Fig. 1. Placement of decoupling capacitors

To ensure effective placement of decoupling capacitors, various software tools are used, including the PrimeRail tool developed by Synopsys [2,3]. PrimeRail gives an opportunity to evaluate the performance of individual parts of the IC based on the performance of the power network, electromigration, dynamic and static current values.

In order to ease calculations and increase the effectiveness of placement of decoupling capacitors, simplified models of individual IC parts are used [4] (Fig. 2).

Considering the VDD voltage equal to 1V, it is possible to calculate the V load voltage as follows [4].

$$I_{load} = \begin{cases} 0, & \text{when } t < t_0 \\ \mu t, & \text{when } t < t_{max} \\ \mu(2t_{max} - t), & \text{when } t < 2t_{max} \\ 0, & \text{if } t > 2t_{max} \end{cases} \quad (2)$$

Where:

t_0 – start time of current variation

t_{max} – maximum current time

t_s – end time of current variation

μ – slope of current variation with time axis

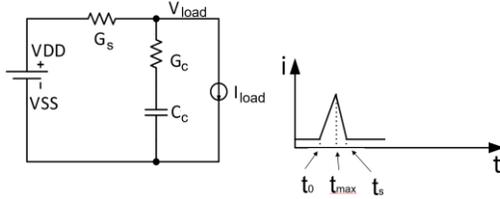


Fig. 2. Simplified model of decoupling

$$\tau = \frac{(G_s + G_c) * C_c}{G_s * C_c} \quad (3)$$

$$V_l = 1 - \frac{\mu}{G_c} \left(t - \frac{C_c}{G_s} (1 - e^{-t/\tau}) \right) \quad (4)$$

Where:

V_l – voltage drop on load

G_s – transconductance of power supply grid

C_c – capacitance of decoupling cap

G_c – active resistance of capacitor

From (2) - (4) it becomes clear that in any part of IC, where the value of the load is equal to the I_{load} and it reaches its maximum value at t_s time, to bring the supply voltage to VDD value, following actions are needed:

1. Increase the capacitance of C_c capacitor,
2. Increase the G_c transconductance value by placing the capacitor as close as possible to the specified part of the IC and by minimizing the resistance between the power rail and capacitor,
3. Increase the power supply's G_s transconductance by decreasing the parasitic resistance of power grid.

II. DECOUPLING CAPACITOR DESIGN IN MODERN CMOS TECHNOLOGY

There are several implementation ways of decoupling capacitances in CMOS technology.

1. Design of decoupling capacitances by using MOS transistors [5]
2. Design of decoupling capacitances by using MOM structure [5]

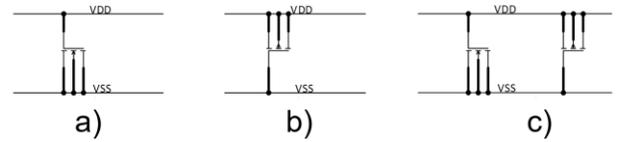


Fig. 3. Decoupling capacitances based on a) NMOS transistor b) PMOS transistor c) CMOS transistors

Capacitance value based on MOS technology is determined by technology process and by device sizes.

$$C = \frac{\epsilon \epsilon_0 W L}{t_{ox}} \quad (5)$$

Where:

ϵ – dielectric permittivity coefficient of SiO_2 between gate and bulk of transistor

t_{ox} – thickness of oxide

w and l – Width and length of MOS device

ϵ_0 – permittivity of free space ($\epsilon_0 = 8,85 \cdot 10^{-12}$ F/m)

From equation (5) it is obvious that for high capacitance value big device sizes are required.

For MOM based structured capacitances comb drive capacitor structures are used. The value of capacitance calculated by following equation [6].

$$C = \frac{\epsilon \epsilon_0 h L}{d_{ox}} \quad (6)$$

Where:

ϵ – dielectric permittivity coefficient of SiO_2 between metal layers in same surface

h – height of metal,

L – length of parallel metal layers,

d_{ox} – length of dielectric which separates metal layers on same level.

To have bigger capacitance values several metal layers are used. For this type of capacitors capacitance calculation is done by (7).

$$C_{tot} = \sum \frac{\epsilon \epsilon_0 h L}{t_{ox}} + \sum \frac{\epsilon \epsilon_0 W L}{h_{ox}} \quad (7)$$

From (7) it is obvious that MOS based capacitors capacitance is more dependent on PVT variations (35%) than MOM based caps (10%). On the other hand, applied voltage higher than the threshold value of transistor, gives an opportunity to have high capacitance value. That's why designers are still using this structured capacitor.

III. PROPOSED METHOD

To have less dependence from PVT variations and occupy less area new method is proposed in this paper.

Considering advantages and disadvantages of described above new method is developed, which idea is to reduce PVT dependence of total capacitance, by implementing both methods semantically. This method gives an opportunity to reduce occupied area (Fig. 4).

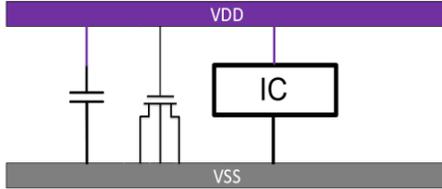


Fig. 4. Decoupling capacitances by usage of MOM and MOS structures combined

To clarify the main idea of this method following calculation was done for 100fF capacitance with 0.3-0.6V voltage applied to it which met during design of analog filters. [6], voltage-controlled oscillators in phase locked loops, etc.

The main idea of proposed method is to compensate the 30fF capacitance lose by using MOM capacitor, because of MOM capacitance is 10 times smaller than MOS's. To get required value proposed method is to use 3 times bigger area than MOS capacitor have. But for effective area usage it is more suitable to use higher metal layers. As showed in equation (7) to have greater capacitance higher metal layers should be used (Fig. 5).

Hence if consider the fact that MOM capacitor ensures 10fF capacitance, then to get 30fF capacitance, design of MOM structured capacitor which area is equal to MOS caps area with 3 metal layers is needed.

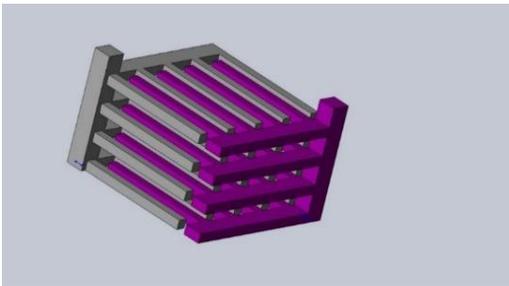


Fig. 5. MOM structured capacitor with additional metal layers

In worst case to have less PVT variation an extra metal layer has been added. As a result, in worst case the total capacitance will be equal to 100fF. Seems that loss will be more than 1.5 times and that makes this method useless. To exclude that statement above it is proposed to place MOM capacitor above MOS capacitor instead of placing them side by side in IC. Such placement is two times more area efficient than placing capacitances side by side (Fig.6)

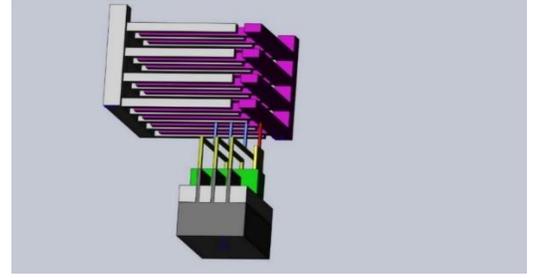


Fig. 6. Parallell placing of MOM and MOS capacitors in layout

IV. RESULTS

The main difficulty during MOS structured capacitor design is capacitance dependence on PVT variations (Fig 7-8).

As higher V_{gs} voltage is, as much amount of charge is accumulated in channel area, therefore capacitance value stands bigger in MOS structured capacitors. This dependence is more expressed in subthreshold voltage region, in case of which number of free carriers in channel area stands lower [7,8]. Process and temperature variations change threshold voltage of transistor, thereby making capacitance value dependence from applied voltage more visible (Fig. 9-10).

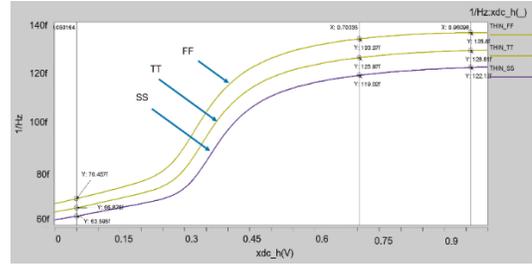


Fig. 7. Thin gate MOS transistors capacitance variation dependence on PV variations

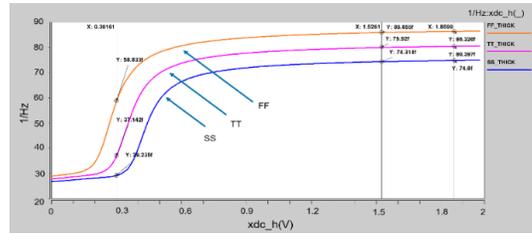


Fig. 8. Thick gate MOS transistors capacitance dependence on PV variations

Compared with MOS capacitors MOM structured capacitors value is more stable from external factors. From fig. 11-13 is visible that capacitance value is stable from voltage and temperature variations, and its value changes from process variations. But it is difficult to get high capacitance value with MOM structure because it occupies big area.

TABLE I. MOM AND MOS STRUCTURED CAPACITORS CAPACITANCE VALUES DEPENDANCE ON PVT VARIATIONS

Parameter name	Temperature (°C)	Voltage(V)
ΔC of MOS (%)	± 10	-25/+10
ΔC of MOM (%)	0,07	0

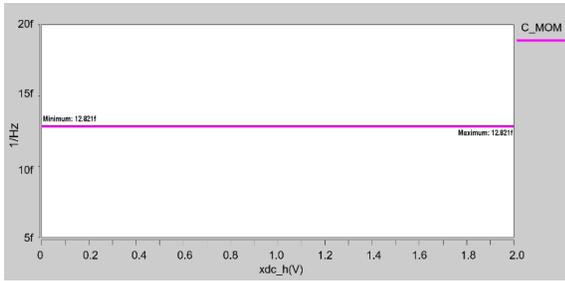


Fig. 9. MOM structured capacitors capacitance dependence on voltage variations

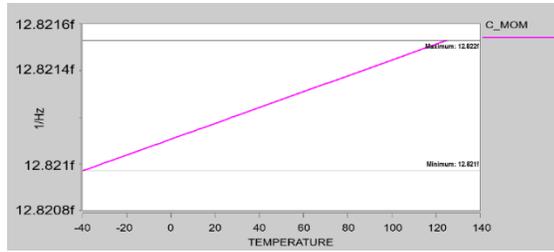


Fig. 10. MOM structured capacitors capacitance dependence from temperature variations

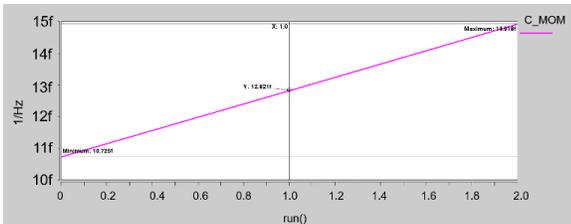


Fig. 11. MOM structured capacitors capacitance dependence from process variations

Calculation of decoupling capacitors capacitance value using Decap Calculator tool.

Decap Calculator software tool has been developed to implement the proposed methods and solutions, automate their implementation, make the IC design easier and less time-consuming. The software tool is based on Synopsys Inc's Galaxy Custom Designer [9] and HSPICE [10] software tools.

Usage of Decap Calculator tool gives an opportunity in early stage of design, when the layout is not fully implemented, values of currents which are obtained from schematic simulation, calculate required capacitance value and area of decoupling capacitor. For capacitance calculation input data is imported in tool, which in future are used in mathematical calculations (Fig. 13).

Input data requires:

1. ".lib" file, which contains technology specific SPICE [10] models of transistors, resistors and other components.

Models also include information about dependence from temperature and applied voltage.

2. Occupied area of capacitor. Tool informs when it is impossible to fit in given area, also it gives information about amount of capacitance and how much space is needed to have required capacitance value using MOS and MOM structures combined.
3. The number of available metal layers for MOM capacitors.
4. The minimum value of applied voltage.
5. Operating frequency (as higher the operating frequency, as bigger noise in supply rails).

Output of the calculation is capacitance value of MOM and MOS capacitors, their occupied area and needed metal layers for MOM capacitor.

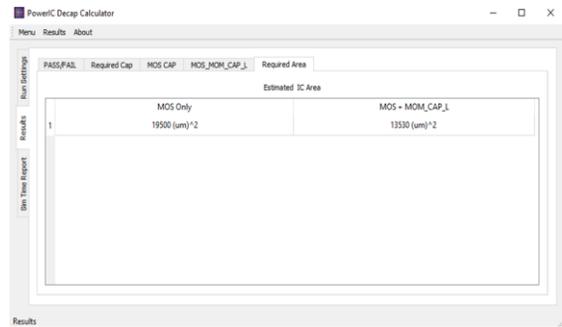


Fig. 12. Required area calculation window of Decap Calculator tool

REFERENCES

- [1] Chia J. Design, layout and placement of on-chip decoupling capacitors in IP blocks. - The University of British Columbia, 2004. - 72p.
- [2] Safaryan K. Power noise optimization with decoupling capacitors 2017 IEEE East-West Design & Test Symposium (EWDTS) Novi Sad, P.1-4.
- [3] PrimeRail User Guide. – Synopsys Inc, 2017.
- [4] Haihua S., Sapatnekar S., Nassif S. Optimal decoupling capacitor sizing and placement for standard-cell layout designs //IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. - April 2003. - Vol. 22, no. 4. - P. 428-436.
- [5] Razavi B. Design of Analog CMOS Integrated Circuits. - Mc Graw Hill India, 2nd edition, 2017. – 782p.
- [6] Charania T., Opal A., Sachdev M. Analysis and Design of On-Chip Decoupling Capacitors //IEEE Transactions on Very Large-Scale Integration (VLSI) Systems. - April 2013. - Vol. 21, no. 4. - P. 648-658.
- [7] Loikkanen M., Rostamovaara J. PSRR improvement technique for amplifiers with Miller capacitor //2006 IEEE International Symposium on Circuits and Systems. - Island of Kos, 2006. – P. 1394-1397.
- [8] Characterizing power delivery systems with on/off-chip voltage regulators for many-core processors /X. Wang, et al //2014 Design, Automation & Test in Europe Conference & Exhibition (DATE). - Dresden, Germany, 2014. - P. 1-4.
- [9] Galaxy Custom Designer Schematic Editor User Guide, Synopsys Inc. - 2014. -236p.
- [10] Hspice Reference Manual, Synopsys Inc.- 2017. -846

Making System Level Test Possible by a Mixed-mode, Multi-level, Integrated Modeling Environment

Nooshin Nosrati¹, Katayoon Basharkhah¹, Rezgar Sadeghi¹, Carna Zivkovic², Christoph Grimm², Zainalabedin Navabi¹

¹School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

²Kaiserslautern University of Technology, Postfach 3049, 67663 Kaiserslautern, Germany

{nosrati.nooshin, basharkhah.kt96, rr.sadeghi, navabi}@ut.ac.ir, {zivkovic, grimm}@cs.uni-kl.de

Abstract— Nowadays electronic systems are moving toward more complex designs with various computation and communication blocks. In addition to test requirements for individual system blocks, the functionality of the overall system must also be tested. Conventional test methods cannot satisfy this requirement due to their limited scope, and time and cost constraints. For this purpose, the concept of system-level test (SLT) has gained attention. However, there are different views on SLT in the literature. Some works consider board-level testing of a complete system as SLT, and others propose system-level fault models, test generation, and design-for-testability for analog, digital, communications, and software parts of a system. Our proposal is a SystemC-based integrated test platform for the SLT model. This model must anticipate the behavior of the system and the effects of different components on each other. Furthermore, it can observe faults on not only the outputs but also intermediate signals of the system.

Keywords—Abstract Model, Mixed Model, System-Level Test (SLT), Integrated Environment, SystemC.

I. INTRODUCTION

Cyber-physical systems (CPSs) consist of a large number of analog and digital parts, sensors, and actuators that are networked and orchestrated by software systems. CPS complexity causes conventional test methods not to be applicable due to test time, cost and fault coverage constraints. In addition, testing of these systems in the context of the hotspot and high stressed environment is crucial because of their use in highly critical applications like automotive [1, 2].

System-level testing (SLT) refers to those test methods applicable to CPSs in an efficient time, cost and fault coverage. There are different views and definitions of SLT in the literature. Many engineers believe that the main function of SLT is to keep IC Defective Parts per Million (DPPM) at a lower level. Some believe that SLT is a too-expensive test step which can be avoided through precise and rigorous fault simulation in low physical and electrical levels [4, 9].

In [5], SLT has been introduced as a platform where a Device-Under-Test (DUT) can be tested functionally using more non-deterministic than deterministic test patterns. This is the opposite of Automatic Test Equipment (ATE) testing where deterministic test patterns are mostly used to test a DUT. Based on [3], SLT is complementary to structural testing. Based on this definition, SLT should be selective and focus on areas of the design that have not been covered by structural testing due to its constraints. Some examples of faults that may not be covered by structural test methods are cross clock domain interactions, power domain switching, resource contention by multiple cores, and faults not covered by fault models, or effects due to the interaction of faults.

We believe that SLT is to test the whole system functionally and not structurally. We consider SLT complementary to structural testing, and not necessarily to be

applied to all parts of the design equally. It can be done for some more and some less based on the application and importance of the components of the system. SLT should cover various components of the design including analog, digital, hardware, software, sensors, actuators, etc.

In order to achieve STL with the above definition, various components of the system need to be modeled individually, and the system model should be formed by the integration of the component models. The level of abstraction of the models depends on criticality and type of the components in the application. The closest model to the real function of the components is more time and cost consuming. Modeling of various components is helpful in fault modeling, test pattern generation and fault simulation.

This paper proposes a SystemC-based platform as an integrated environment to model various types of system components. Using SystemC and its extensions like SystemC-AMS and TLM, we are able to model analog, digital, hardware, software, and communication parts of a system in a single description language. These components can be modeled in different abstraction levels. Using such an integrated environment, an SLT platform can be used for fault modeling, test pattern generation, fault simulation, verification, and design for testability. At the same time, interaction of components with each other and the stability of the whole design are considered. As a case study, this paper discusses how to model, test and translate electrical low-level deviations and faults of analog and communication components, and studies their impact on system properties.

On the other hand, SystemC verification provides key facilities to construct advanced reusable, automatic and organized verification. SystemC verification facilitates data introspection, randomization and seed management and reproducibility of a simulation run, modular constrained randomization, weighted randomization, transaction monitoring, and recording, and sparse array support which are essential for testing procedures.

The rest of the paper is as follows: Section II reviews the present definitions of SLT and solutions presented to address the fault coverage, time and cost constraints. Section III discusses what is required of SLT, and concludes that modeling a system is the main requirement for SLT to be defined for the system-level test in the design process, and after implementation of the system. Section IV proposes the SystemC hardware description language and its derivatives for system-level modeling. It shows how various test applications can benefit from a SystemC system-level model. Section V concludes the paper.

II. LITERATURE REVIEW

The system-level test is a new term that refers to testing cyber-physical systems and, in general, systems containing

hardware, software, communication and, multiple chips in a package, in actual use circumstances.

Various works in the literature have different definitions for SLT and they present their approaches based on a certain interpretation of this concept. In general, they all share the fact that SLT considers an integrated system and one whose components are a mix of analog, digital, software, with varying types of interconnections. Some of these works are presented in the paragraphs that follow.

Some works [9] consider SLT in board-level (post-manufacturing). Board-level SLTs are applied to capture functional faults. In these works, it is suggested to replace the final test of the system on automatic test equipment (ATE) with SLT. Board-level SLT is too costly and has limited throughput. Work in [6] addresses these issues by proposing a stressed structural test method. This method selects partial chips to apply system-level tests only on them. For this purpose, support vector machine (SVM) as a classifier uses stressed test data to select a subset of total chips. Another work [7] filters a portion of chips from the system-level test by data analytic techniques from upstream ATE test prediction.

All of these works perform SLT on a post-manufactured system where there is no chance to apply remedies for improving the performance and reliability of the system. On the other hand, determining whether the fault is caused by software or hardware is a challenging work for engineers. System-level design-for-testability (SL-DFT) can address these issues by modeling and evaluating the system in the design phase. SL-DFT provides solutions by using controllability and observability [3].

In contrast to [9], work [5] considers SLT platform besides ATE. This work defines SLT as several test cases on operating systems under various test conditions. In this paper, instead of a sequential SLT, mutually exclusive test cases are executed concurrently. This concept is called Concurrent System-Level Test (CSLT). The proposed method helps in reducing the test time and also simulates end-user experience more effectively. However, this paper only focuses on operating system test while SLT also encompasses hardware components including processor cores, accelerators, memory blocks and communication infrastructures.

On the other side, [10] is a comprehensive review on SLT that focuses on hardware components, especially cores embedded in a system on chip (SOC). This work reviews different SLT requirements including fault model, test generation, DFT and system-level dependability analysis. Regarding SL-DFT, it proposes a hybrid built-in self-test (BIST) that improves fault coverage by combining pseudo-random test patterns with deterministic ones. This also results in the test time reduction.

Our contribution is to cope with the aforementioned issues. This work suggests a mixed-mode SLT platform which considers both hardware and software aspects of a system. This platform provides an integrated environment for system-level design and test (SLDT) in multi-level.

III. SLT REQUIREMENTS

Looking back at the digital design history pages, we see that SLT is not a new concept in the test area, just the systems we are testing are different. In different eras, the “system” has referred to different levels of hardware abstraction. A few years ago, for its time, gate-level designs we considered as

complete systems, and testing gate-level circuits was considered as a system-level test, although syntactically not referred to as such. Looking at SLT from this point of view, we reach a set of mandatory requirements that must be met for testing at any level. The difference is in today’s systems, where integration, multi-mode systems, and multi-level descriptions make testing more difficult.

A. Need for a model

An important issue at any level of test is a good and uniform fault model at that level. To be effective, a fault-model requires a model of a system to treat the fault with. Some works have discussed system-level fault model (SL-FM), system-level test generation methods (SL-TG), and system-level design-for-testability (SL-DFT) without considering any SLT model. In a way, without a system-level comprehensive model to act as the base of all test methods, you are putting the carriage before the horse.

B. Golden model

The need for having a model for SLT begins with the requirement of having the good behavior of a system or its golden model. With an integrated model, the golden model considers the overall functionality of various components and their effects on each other. This model not only provides the correct behavior of the overall system, but also the good functionality of the interfaces of various components that may be analog, digital or software. Such a model can be used to analyze the post manufactured system and verify its operation against the correct values obtained by simulation of the golden model.

In addition, a comprehensive integrated model in which all modules are described separately and their interfaces are clearly defined can be analyzed for faults and the effects of faults on the interfaces of various modules as well as the outputs.

Having a comprehensive model of a manufactured system facilitates post-manufacturing SLT that is the focus of most of today’s work on SLT. Although the post-manufacturing test is not the focus of this paper, using the same modeling strategy for post-manufacturing SLT, and for the design process, which is the focus of this paper, eases the former and facilitates the latter.

C. Mixed-mode model

A mixed-mode model of a system is one in which analog, digital, sensors, actuators, and software parts are described in their own specific way of modeling, and they interact with each other just like the actual physical parts do.

Having a comprehensive mixed-mode model in the design process provides the opportunity of fault injection and evaluation of fault-tolerant methods that we insert in our system. Figure 1 shows how an SLT model is used in this scenario. As shown, the system is modeled in an integrated environment (lower left box), then faults can be inserted in various parts of the system using the specific fault models for that part of the system (upper left two boxes). As shown, MDSI faults are injected for the interconnects, Affine Arithmetic uncertainties are used for faults in the analog components, stuck-at faults and other more abstract functional faults are used for the digital parts of a system, and instruction-level faults are used for the software parts of the system.

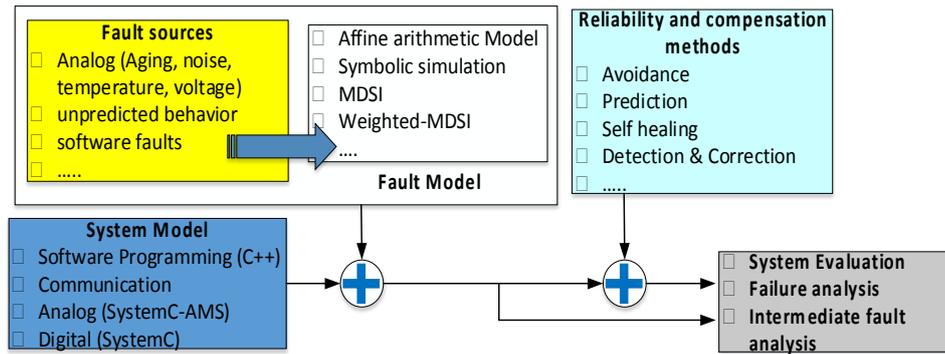


Figure 1: A flow of SLT model.

After evaluating good and faulty behaviors of a system, reliability methods can be applied to make the system reliable. As shown in Figure 1, reliability methods such as fault avoidance techniques, prediction, and self-healing methods (upper right box) are inserted in hardware or software and modeled properly to be evaluated for their effectiveness. The system model with reliability methods can be evaluated and analyzed for failures, performance degradation, and other types of system malfunctionings. Based on such analysis and perhaps considering overall power and other physical constraints, a proper reliability technique will be chosen for our system.

Another place where a mixed-mode model that includes software, analog, and digital parts becomes useful is in considering faults and compensation of faults as a whole. A fault or imprecision of a sensor can be partially or fully compensated by its digital interface circuitry or the software that interprets its value.

Due to the paramount importance of software in nowadays systems, not only most functions can be implemented by software, but also mechanisms for error detection and error reactions. The ability to study the interaction between faults in all components and in the system-level software is perhaps the most important factor to be considered for modeling of faults.

D. Multi-level model

A multi-level modeling environment is one that allows the model of a specific component to be described at various levels of abstractions. For example, a gate-level circuit can be described at the gate-level, R level, or pure functional without any concern about signals and/or registers.

Ability to have a multi-level modeling tool becomes important when the top-down design of a system is considered. Going back to the time that systems were no bigger than several RTL components put together, the top-down design process of such systems included testing of the components as the design evolved. Although in the early times of RT level design, design and test were done separately, and the design engineer and the test engineer had to go to war for chip real estate, RTL design eventually evolved to one pass of Design-and-Test, or DfT (Design for Test). This success was partially due to the evolution of hardware description languages, design tools based on HDLs, and the fact that at any point of design an overall HDL model of the system was available, even though some parts were still in their abstract form and not implemented as hardware units. Multi-level

HDLs, where one part was described at the functional level, while another at the gate or RT level was an important enabling factor here.

We expect the design of a system at the system-level to evolve in a similar fashion to RTL. This means that in addition to requirements discussed in the previous subsections, a multi-level modeling platform is also necessary for SLT. Our modeling tool that can be used for System Level Design and Test (SLDT) must be a multi-level HDL

We have discussed several topics that we consider as requirements for a modeling language that can be used for SLT. This includes having a functional model for good behavior, ability to control and observe component boundaries, mixed-mode, and multi-level modeling. We propose our modeling language in the next section.

IV. CONVERGE IN UPON A MODELING LANGUAGE

There is a need for an integrated environment to model a system including digital, analog, software programs and inter- and intra-block communications in which it can fulfill all above features. SystemC-based modeling is suggested here.

A. The language

The SystemC hardware description language and its derivatives including SystemC-AMS (analog and mixed-signal), SystemC-Verification (SCV), TLM-1 and TLM-2.0 are C++ class libraries for modeling and verification.

The main SystemC core focuses on timing and concurrency that are the center point of any hardware description language. SystemC allows concurrent simulation of digital, analog, software, and interfaces that connect these parts together. With this concurrency at the center point, AMS extension of SystemC provides a very convenient interface for modeling analog electronics and non-electric modules. Then again, concurrent with all other parts of a system, the C++ base of SystemC gives it an easy interface with software programs for describing programs that run on a system or for ISS (instruction set simulation) of software parts of a system.

TLM-1 and TLM-2.0 also handle communications between processing elements by general-purpose predefined utility channels and memory interfacing, respectively. These facilities hide handshaking and data handling in abstract functions. This way, the designer can focus on the functionality of a system, rather than communication details. SCV provides facilities to simplify test procedure including fault injection, random test, and transaction recording.

SystemC uses two different simulation engines; one for the analog parts and one for the rest. Continuous simulation (CS) is used for running the AMS extension, while it uses a discrete-event simulation (DES) engine for the rest of the system. Hence, inter-block communication between CS and DES becomes necessary, that is done by predefined converters. Meanwhile, inter-communications between DES blocks and intra-block communications are done easily by signals and variables due to using the same simulation engine.

Describing various components of a system at various levels of abstraction is facilitated in SystemC for analog and digital parts alike. A digital component can be described using its gate level structure, its RT level description, a software program describing its functionality, or any form of description in the middle. Likely, an analog part of a system can be described using its passive elements, the mathematical expression it implements, or a program-like description of the component.

B. Utilization of SystemC

There are different abstraction levels of a system that range from describing a component by its physical components, to a behavioral description of a module. Likely, faults in a module can correspond to the individual physical elements of a module or to the way it behaves. Different blocks of a system can be modeled based on how the physical parameters of the blocks are to be considered. This consideration is determined based on the specific application.

As discussed, we have to think about a model before we take on SL fault modeling, SL test generation, etc. SystemC sounds an appropriate integrated environment to model a complex system including digital and, analog, hardware with inter- and intra-block communications at one hand, and software functions at the other hand. However, for test applications, modeling faults is required that because of varying nature of components that are used in a system (i.e., digital, analog, software), a unique fault model cannot be reached. Therefore, for performing SLT, specific fault models of each block must be inserted individually while evaluating the functionality of the overall system.

C. A case study using SystemC

In order to illustrate how SystemC and its derivatives can be useful in modeling a system so that SLT can be used for it, a small communication system is used here as a case study.

Figure 2 shows the block diagram of our case study system. This is a software-defined radio system (SDR) that filters the data from its antenna and delivers it back to the antenna for transmission. The system has several digital circuits, analog circuits, an antenna, and ADC and DAC converters. There is a communication channel for data reception and transmission. In SDR, radio components including mixers, filters, modulators/demodulators, and detection circuits are implemented in a programmable medium to provide increased flexibility and capabilities. The programmable parts are the digital and software components. In our case study, we place fault detection and avoidance methods in the digital and software programmable parts.

To examine our aforementioned reliability techniques, we start with an integrated system model in which digital, analog, and software components of the system are modeled according to the type of hardware they represent. The modeling is done with a SystemC-based platform.

There are comprehensive works on system-level testing of digital blocks in [10]. We focus on the analog and communication channel modeling to make the examination flow described in Figure 1 possible. This means each needs a model, a fault model and a technique for making it reliable.

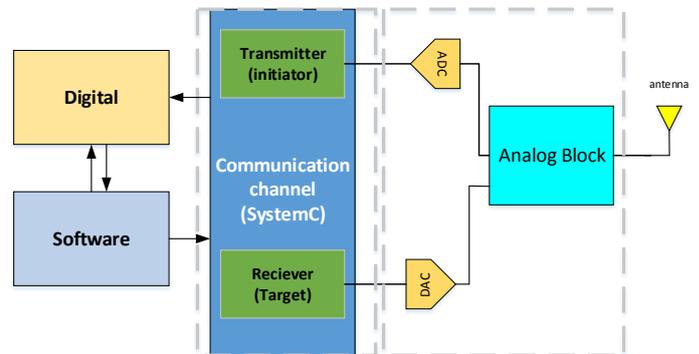


Figure 2. A case study system.

1) *Analog Circuit.* The analog circuit that we are investigating for reliability as part of our SLT is an amplifier. We will talk about the system model, the way we are faulting it, and the way we intend to compensate for the fault.

a) *System model.* This analog circuit includes a low-noise amplifier (LNA) block, a variable-gain amplifier (VGA) block, and a few other analog discrete parts. These blocks are modeled using SystemC-AMS. For this modeling, Affine arithmetic (AA) can be used that properly models deviations in value and inaccuracies in components or variables of an analog circuit. Affine Arithmetic (AA) is range arithmetic that overcomes the error explosion problem of Interval Arithmetic. AA keeps track of correlations between quantities represented as ranges [11]. This procedure is used in [13]. In SystemC-AMS the behavior of the system is described using data type AAF for the input and output ports. To model the LNA, it is assumed that the exact gain value is not known. So a range of values is assigned to the signal values, and the computations are based on the midpoint of the range and maximum deviation from the midpoint.

b) *Fault injection.* For modeling analog faults, they are classified into different kinds of fault sources like parametric faults e.g. due process, voltage and temperature variations, aging, but as well unforeseen defects that lead to non-deterministic behavior. This classification is based on the mathematical modeling background of the faults, i.e. range or probabilistic methods. This method computes the worst-case behavior over the considered ranges of input and initial operating conditions. It saves the sequence of inputs or parameters that lead to this worst case. Then, based on the previous and current conditions, the solver decides on the new relational condition and uncertainties.

c) *Self-compensating.* Parametric faults are mostly compensated by built-in robustness of the system, e.g., by control loops at various levels or even in system-level software. The fault model is simulated with a symbolic simulation at the system level. During the simulation run, all inputs and corresponding output affecting the behavior of the circuit are tagged. Once there is a failure, the system is rewound to the last correct state.

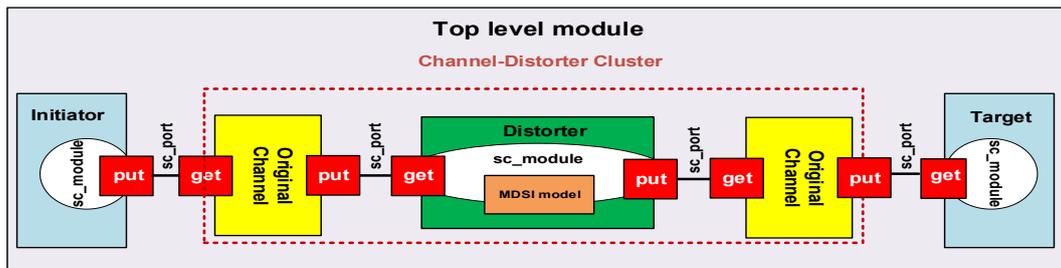


Figure 3. Fault injection in the communication channel.

2) *Intra-digital block communication.* Another part of the system of Figure 2 that we are analyzing for fault injection and examining its reliability is a bus that runs between several of the systems digital parts. The paper proposed in [12] has modeled a relatively complex communication channel for abstract initiator-target communications using the SystemC language. Then, a mechanism for inserting fault models into the communication lines is proposed. Finally, they have evaluated a faulty communication channel in the presence of various reliability methods including fault avoidance, detection and correction techniques. We show that SystemC models provide a mechanism for fault injection as well as evaluation of ways of circumventing faults.

a) *System model.* ESL communications are described in SystemC as channels. Although there are specialized channels in SystemC for specific cases, e.g., TLM-2.0 for processor-memory communications, work in [12] takes a more general approach, and uses the channel definition utility of SystemC. The channels are designed to be able to inject various interconnect faults, e.g., crosstalk and noise. Details of modeling communication channels is addressed in [12].

b) *Fault injection.* Fault injectors are presented as separate SystemC modules that can be instantiated within the channels. The proposed mechanism for fault injection handled by putting a distorter module is shown in Figure 3. This module, that is sandwiched between two identical copies of the original channel, distorts the data transferring through the channel based on a given fault model. The work in [12] offers two fault models that can be inserted in the distorter module. These models are an abstract MDSI [14], and a more detailed model that is referred to as Weighted-MDSI [8]. Both these models are described in SystemC. The noise insertion method of this paper can also use a physical RC/RLC network module described in SystemC-AMS.

c) *Reliability methods.* The paper in [12] has developed several reliability methods including shielding, duplication, parity, and Hamming for making channels fault-tolerant and testable. For this purpose, some redundancy hardware and also detection and/or correction hardware have been added to the channel interface of the initiator and to the interface of the target, respectively. The modeling discussed here allows evaluation of these error-correcting methods for their effectiveness, hardware overhead, and power consumption.

V. CONCLUSION

To cope with many challenges faced by testing complex systems with various computation and communication blocks, SLT is considered. SLT requires an integrated environment for mixed-level, modular, functional, and multi-level

modeling of a complex system. A model for this system is to be used for test generation, reliability analysis, and system-level DFT. The center of all such requirements is fault models that can describe faults of components of a system according to the nature of the component. The SLT model must satisfy requirements including mimicking system's good behavior, analysis effects of different components on each other, observing faults on the interfaces of various components, as well as on the outputs of the system. In this paper, SystemC is suggested as an appropriate integrated environment to fulfill all mandatory requirements for a model to be considered appropriate for the base of SLT. A case study demonstrates how a SystemC model of a system can be used for SLT.

REFERENCES

- [1] <https://www.advantest.com/system-level-test-systems>
- [2] <https://semiengineering.com/toward-system-level-test>
- [3] H. H. Chen, "Beyond structural test, the rising need for system-level test," 2018 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), Hsinchu, 2018, pp. 1-4.
- [4] J. Kim, S. Chon, and J. Park, "Suggestion of Testing Method for Industrial Level Cyber-Physical System in Complex Environment," 2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), Xi'an, China, 2019, pp. 148-152.
- [5] D. K. R. Tipparthi and K. K. Kumar, "Concurrent system level test (CSLT) methodology for complex system-on-chip," 2014 IEEE 16th Electronics Packaging Technology Conference (EPTC), Singapore, 2014, pp. 196-199.
- [6] J. J. Liou, M. T. Hsieh, J. F. Cherng and H. H. Chen, "Cost reduction of system-level tests with stressed structural tests and SVM," In 2015 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), 2015, pp. 177-182.
- [7] H. H. Chen, "Data analytics to aid detection of marginal defects in system-level test," In 2016 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), 2016, pp. 1-4.
- [8] R. Sadeghi, N. Nosrati, K. Basharkhah, and Z. Navabi, "Back-annotation of Interconnect Physical Properties for System-Level Crosstalk Modeling," in Test Symposium (ETS), 2019 24th IEEE European, 2019.
- [9] S. Biswas and B. Cory, "An Industrial Study of System-Level Test," in IEEE Design & Test of Computers, vol. 29, no. 1, pp. 19-27, Feb. 2012.
- [10] M. Sonza Reorda Z. Peng, and M. Violante, eds, "System-level Test and Validation of Hardware/Software Systems," Vol. 17. Springer Science & Business Media, 2006.
- [11] C. Radojicic, T. Purusothaman, and C. Grimm. "Symbolic Simulation of Mixed-Signal Systems with Extended Affine Arithmetic," In edaWorkshop Proceedings. VDE Verlag, May 2015.
- [12] N. Nosrati, K. Basharkhah, R. Sadeghi, and Z. Navabi, "An ESL Environment for Modeling Electrical Interconnect Faults," in 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2019.
- [13] C. Radojicic, C. Grimm, J. Moreno, and X. Pan. "Semi-symbolic analysis of mixed-signal systems including discontinuities," In Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014. 2014, pp. 1-4.
- [14] S. Chun, Y. Kim, and S. Kang, "MDSI: Signal integrity interconnect fault modeling and testing for SOCs," J. Electron. Test., vol. 23, no. 4, pp. 357-362, 2007.

Fast and Efficient Implementation of Lightweight Crypto Algorithm PRESENT on FPGA through Processor Instruction Set Extension

Abdullah Varici
EE Dept.
Ozyegin University
Istanbul, Turkey
abdullah.varici@ozu.edu.tr

Gurol Saglam
CS Dept.
Ozyegin University
Istanbul, Turkey
gurol.saglam@ozu.edu.tr

Seckin Ipek
EE Dept.
Ozyegin University
Istanbul, Turkey
seckin.ipek@ozu.edu.tr

Abdullah Yildiz
CSE Dept.
Yeditepe University
Istanbul, Turkey
ayildiz@cse.yeditepe.edu.tr

Sezer Gören
CSE Dept.
Yeditepe University
Istanbul, Turkey
sgoren@cse.yeditepe.edu.tr

Aydin Aysu
ECE Dept.
North Carolina State University
Raleigh, NC, USA
aaysu@ncsu.edu

Deniz Iskender
CS Dept.
Ozyegin University
Istanbul, Turkey
deniz.iskender@ozu.edu.tr

T. Baris Aktemur
CS Dept.
Ozyegin University
Istanbul, Turkey
aktemur@gmail.com

H. Fatih Ugurdag
EE Dept.
Ozyegin University
Istanbul, Turkey
fatih.ugurdag@ozyegin.edu.tr

Abstract—As Internet of Things (IoT) technology becomes widespread, the importance of information security increases. PRESENT algorithm is a major lightweight symmetric-key encryption algorithm for IoT devices. Compared to the Advanced Encryption Standard (AES), PRESENT uses a lower amount of resources while achieving the same level of security. In this paper, we implement PRESENT with different design methodologies including hand-coded RTL, Vivado HLS, PicoBlaze, VerySimpleCPU (VSCPU) based microcontrollers, and a customized VSCPU. The customized VSCPU design is based on optimizing the instruction set architecture for the algorithm specifics of PRESENT. Our results show that the customized VSCPU design methodology can be more efficient than HLS and PicoBlaze while providing the flexibility compared to RTL designs.

Keywords—Internet of Things, Information Security, Cryptography, PRESENT Algorithm, FPGA, VerySimpleCPU

I. INTRODUCTION

The Internet of Things (IoT) is a platform where smart home appliances, various sensors, wearable and many other devices communicate with each other over the Internet. The smart devices surrounding us gather information from the environment and they share the information by using processors and communication units with other smart devices over IoT platform to produce solutions to daily life problems. These smart devices are becoming more and more widely used in various fields such as agriculture, health, safety, education, and urbanism [8]. Such a situation makes information security in these devices covering many areas of our lives an important issue.

Data encryption before transfer and storage is a typical method to prevent unwanted access to confidential information. Encryption algorithms such as Advanced Encryption Standard (AES) [1] and Data Encryption Standard (DES) [2] are widely used for this purpose. However, since these algorithms require a large amount of hardware resources and high power consumption, it is not possible to use them in edge/IoT devices with limited resources. The Lightweight Cryptography field [9] has emerged as a solution to this problem by decreasing processing complexity and the area usage of the chips while providing a reasonable amount of security.

One of the most important components of modern electronic systems are the processors and the software enables the processors to perform the desired operations in order. The instruction sets of many commercially available processors are non-modifiable. This leads to the fact that the instructions, which the applications sometimes do not use at all, consume unnecessary hardware resources. In addition, much needed application specific instructions are not in the instruction set causing serious decreases in performance or the application cannot perform the desired operation at all. To avoid these problems, there is a need for application-specific hardware design and changing the instruction set of processors.

In this work, we discuss a methodology by extending the instruction set of the VerySimpleCPU (VSCPU) [5] for efficient design of any application-specific algorithm on the devices with limited resources. To that end, we implement and compare the results obtained by designing the desired algorithm with pure RTL coding, Vivado High Level Synthesis (HLS), PicoBlaze, and the unmodified VSCPU. We use

FPGAs as the target platform for our implementations and choose PRESENT [3], the popular lightweight cryptography algorithm, as the example algorithm in this work.

Our results shows that customizing the VSCPU core can lead to a very desirable solution by combining the best of both worlds. On the one hand, it can provide the flexibility of a software-based solution. On the other hand, it can lead to a more efficient design than HLS and Picoblaze microcontrollers through hardware customization. This work therefore shows an exciting opportunity for IP design of next-generation cryptography standards.

II. PRESENT ALGORITHM

The PRESENT [3] encryption algorithm is a lightweight symmetric-key block cipher algorithm. It has the Substitution Permutation Network (SPN) structure and has a 64-bit block length along with an 80-bit or 128-bit key length. As shown in Fig. 1, the algorithm consists of 31 rounds and the last round of key addition. A round consists of the following 3 functions: add round key, substitution box layer, and permutation layer.

Adding the round key is simply the XOR operation of the state data block with the round key. The nonlinear substitution layer uses identical Substitution Box (SBox) that converts a 4-bit input to a 4-bit output. Fig. 2 shows the input-output relationship of a single Sbox. In the permutation layer, the bit value in the index m is moved to the index $n = P(m)$ for all indexes. Fig. 3 provides the construction of the P function.

PRESENT uses distinct round-keys at each round. Fig. 4 formulates the key schedule operation, which first performs a 61-bit left rotation. Then, the left-most 4-bit for 80-bit keys and the left-most 8-bit for 128-bit keys are passed through the Sboxes. Finally, the round counter value is subjected to XOR operation with the bits of 19:15 indexes of the previous result. The most significant 64-bit of the value forms the round key.

III. THE IMPLEMENTATION OF THE PRESENT ALGORITHM

To implement the PRESENT algorithm, we use Xilinx's ISE 14.7 software for synthesis, simulation, and placement and routing, and Vivado HLS 2017.1 software for high-level synthesis (HLS). We select the Xilinx XC7A100T-1CSG324 FPGA and use the 80-bit key version of PRESENT. Tables I and II summarize the results.

A. Implementation with Hand-Coded RTL

Designing a custom chip is the first option that comes to mind when designing hardware with low resource consumption and high-performance. However, because of the high cost per product of chip design for small quantities of products, Field Programmable Gate Arrays (FPGA) are preferred. FPGAs are also used to ensure shorter design time and prototyping of ASICs. FPGA design is made using a Hardware Description Language (HDL). The two most commonly used languages are Verilog and VHDL. FPGAs are programmed after the design's synthesis and implementation stages with selected FPGA software.

```

generateRoundKeys()
for i = 1 to 31 do
    addRoundKey(STATE, Ki)
    sBoxLayer(STATE)
    pLayer(STATE)
end for
addRoundKey(STATE, K32)

```

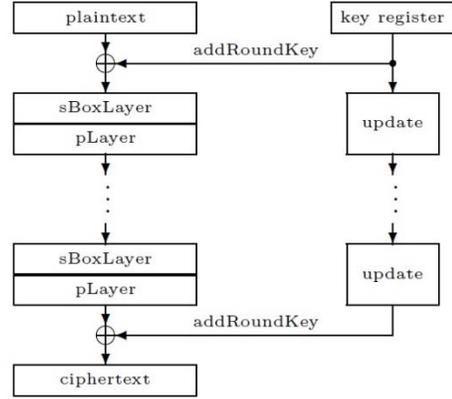


Fig. 1. PRESENT encryption algorithm [3]

x	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
$S[x]$	C	5	6	B	9	0	A	D	3	E	F	8	4	7	1	2

Fig. 2. Substitution box layer [3]

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$P(i)$	0	16	32	48	1	17	33	49	2	18	34	50	3	19	35	51
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
$P(i)$	4	20	36	52	5	21	37	53	6	22	38	54	7	23	39	55
i	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
$P(i)$	8	24	40	56	9	25	41	57	10	26	42	58	11	27	43	59
i	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
$P(i)$	12	28	44	60	13	29	45	61	14	30	46	62	15	31	47	63

Fig. 3. Permutation layer [3]

- $[k_{79}k_{78} \dots k_1k_0] = [k_{18}k_{17} \dots k_{20}k_{19}]$
- $[k_{79}k_{78}k_{77}k_{76}] = S[k_{79}k_{78}k_{77}k_{76}]$
- $[k_{19}k_{18}k_{17}k_{16}k_{15}] = [k_{19}k_{18}k_{17}k_{16}k_{15}] \oplus \text{round_counter}$

Fig. 4. Key schedule [3]

Before carrying out other implementations, we wanted to find out hardware resource consumption and typical performance of the PRESENT algorithm with hard-coded, i.e., pure RTL design on FPGA. For this purpose, we reviewed the RTL design shared on GitHub [4]. We also simulated, synthesized and implemented the design on the selected FPGA without the need for optimization. The design consumed 213 LUTs and 213 FFs from FPGA resources after implementation. It also can work with a clock frequency up to 368 MHz, and it can compute a result in 32 cycles (latency).

TABLE I. MEMORY RELATED INFORMATION ON PROCESSOR BASED IMPLEMENTATIONS

<i>Processor / Memory</i>	<i>PicoBlaze</i>	<i>VSCPU compiled code</i>	<i>VSCPU-80</i>	<i>VSCPU-16</i>	<i>[13]</i>
Instruction Bit Length	18	32	10	12	8
Number of Instructions in the Program	509	2210	16	45	156
Data Bit Length	8	32	80	16	8
Number of Data in the Program	-	58	8	16	48

TABLE II. HARDWARE RESOURCE CONSUMPTION AND PERFORMANCE INFORMATION OF IMPLEMENTATIONS ON FPGA

<i>Implementation / Consumption and Performance Information</i>	<i>Pure RTL</i>	<i>Vivado HLS</i>	<i>PicoBlaze</i>	<i>VSCPU compiled code</i>	<i>VSCPU-80</i>	<i>VSCPU-16</i>	<i>[11]</i>	<i>[12]</i>	<i>[13]</i>
Look-Up Tables (LUT)	213	466	292	8966	1023	270	222	226	377
Flip Flops (FF)	213	765	104	3203	507	291	201	89	N/A
Max. Clock Frequency in MHz	368	125	185	67	141	153	236	172	542
Latency in Cycles	32	29k	360k	558k	502	1402	295	132	20k
Latency	87 ns	233 us	2 ms	8 ms	4 us	9 us	1 us	767 ns	37 us
Latency x (LUTs + FFs)	37k	286M	768M	101G	5.4M	5.1M	528k	242k	14M

With the pure RTL design, the best results were obtained in terms of performance and area consumption and these results were used for comparison with other implementations. The design to be done in this way has disadvantages such that the design is coded and validated manually so that the design time is too long and that the design does not offer any flexibility to the lack of programmability.

B. Implementation with Vivado HLS

Vivado is a complete software that allows performing RTL design, simulation, synthesis, implementation and Xilinx FPGA programming. To make FPGA programming simpler and reduce the design time, Xilinx has developed the Vivado HLS. Thanks to this program, a C-based code can be transformed into a Verilog or VHDL design which allows synthesis and implementation on FPGAs.

We used the C-based design of the PRESENT algorithm from FELIX project [10] on Vivado HLS 2017 and confirmed that the design works correctly. Afterward, synthesis and implementation of the obtained RTL design from HLS were performed. The resulting design consumes 466 LUTs and 765 FFs, achieves a clock frequency of up to 125 MHz, and takes 29124 cycles to complete an encryption operation.

Designing with HLS significantly reduces design time. However, when the obtained results are examined, it is seen that resource consumption increases approximately twice as compared to the pure RTL implementation. In addition, an encryption operation results in 32 cycles in the pure RTL implementation, while 29124 cycles in the Vivado HLS implementation. Considering the maximum clock frequencies that the designs can operate, it is seen that there is a decrease of approximately 2681 times in performance. Moreover, as with the pure RTL design, the design produced by Vivado HLS

does not include a processor, so it does not allow software development.

C. Implementation with PicoBlaze

PicoBlaze [7] is an 8-bit processor with low area consumption and low performance for Xilinx's own FPGAs. This processor goes through the synthesis and implementation stages and works on the resources in Xilinx FPGAs. There are a total of 70 commands in PicoBlaze and all commands are executed in 2 cycles.

In order to implement PRESENT algorithm with PicoBlaze, the algorithm was written using Assembly language and PicoBlaze instruction set. Initially, the Assembly program consisted of 1446 instructions, and the design's program memory would consume a significant amount of resources. To optimize the code, we review it and reduce the number of assembly instructions to 509. Since FFs are used as the memory element in all implementations and in order to make the comparison under equal conditions, the memory elements of PicoBlaze are changed to FFs and LUTs from Block RAMs. Afterwards, we performed synthesis and implementation using the memory file from Assembly code and PicoBlaze's source code for Xilinx ISE. Eventually, the design working with a clock frequency up to 185 MHz consumed 292 LUTs and 104 FFs of FPGA resources and the latency for an encryption was 359943 cycles.

The results show us that, by using PicoBlaze, it is possible to create a low resource consuming design, as in pure RTL design. In addition, due to its processor based architecture, software development on this implementation is possible. However, when the clock frequencies and the number of cycles required for an encryption are considered, PicoBlaze

implementation is approximately 22324 times slower than the pure RTL design.

D. Implementation with VSCPU

VSCPU [5] is a simple and customizable 32-bit processor that can be implemented on FPGAs, with an instruction set simulator, Assembly code generator, and C compiler. VSCPU has 16 instructions in its instruction set and supports unsigned integer arithmetic operations. The RTL codes of the processor are written in Verilog language and the structure can be changed completely as required.

We compiled the C code of the PRESENT algorithm with the C compiler of the VSCPU and extracted the Assembly code. As a result, 2210 32-bit instruction memory, which is stored in a ROM, and 58 32-bit data memory, which stored in a RAM, were needed. The result of encryption operation is computed at the end of 557997 cycles and the design can work with clock frequency of up to 67 MHz using 8966 LUTs and 3203 FFs.

Considering the resulting resource consumption and performance, we can say that VSCPU implementation has the worst results among other options. The reasons for this may be that the instruction set and architecture of VSCPU are not well suited to the PRESENT algorithm, and that the C code compiler does not produce good results. However, this implementation has the advantages that it provides a much faster design opportunity than a pure RTL implementation and allows for software development due to its processor based structure.

E. Implementation with Instruction Set Modified VSCPU

Due to the lack of XOR operation and algorithm specific commands such as SBOX in the VSCPU's instruction set, the compiled Assembly code consumed a lot of memory and took too long to compute an encryption operation. Therefore, to improve performance and reduce resource consumption, we changed the instruction set of the VSCPU and wrote the Assembly code manually.

Firstly, because the key length is 80-bit, we increased the length of the data bus and the words in the data memory to 80-bit length. We also added four new instructions in the algorithm that implement substitution (SBOX), left rotation (RRL), permutation (PER), and XOR functions. We removed instructions that were not used in the algorithm from the instruction set.

Another change in the VSCPU was the transition from a combined memory unit in the Von Neumann architecture to the separate storage of data and instructions, as in the Harvard architecture. The aim was to take advantage of the fact that the bit length of the instructions is shorter than the bit length of the data and to consume fewer resources.

This design required 16 instruction words of 10-bit length and 8 data words of 80-bit length. The design, which completes an encryption in 502 cycles, can work with up to 141 MHz clock frequency, using 1023 LUTs and 507 FFs. When the design obtained is evaluated in terms of performance, it provides much higher performance than other realizations

except pure RTL implementation. However, it gives better results in terms of resource consumption than only VSCPU.

TABLE III. INSTRUCTION SET OF THE VSCPU-16

<i>Instruction</i>	<i>Functionality</i>
XOR A B	$*A = *A \wedge *B$
CP A B	$*A = *B$
ADD A B	$*A = *A + *B$
BZJ A B	if(*B == 0) jump *A
GETW A B	if(B == 0) W = 0 else *A = W; W = 0
SRW3 A B	$*A = (*A \gg 3) W$; $W = *A \ll 13$
SBOX16 A B	if(*B == 0) *A = { sbox(*A[15:12]), *A[11:0] } else *A = { sbox(*A[15:12]), sbox(*A[11:8]), sbox(*A[7:4]), sbox(*A[3:0]) }
PER3 A B	$W = \{ *A[15], *A[11], *A[7], *A[3], *B[15], *B[11], *B[7], *B[3], W[15:8] \}$
PER2 A B	$W = \{ *A[14], *A[10], *A[6], *A[2], *B[14], *B[10], *B[6], *B[2], W[15:8] \}$
PER1 A B	$W = \{ *A[13], *A[9], *A[5], *A[1], *B[13], *B[9], *B[5], *B[1], W[15:8] \}$
PER0 A B	$W = \{ *A[12], *A[8], *A[4], *A[0], *B[12], *B[8], *B[4], *B[0], W[15:8] \}$
BZJi A B	Jump *A+B

To further reduce the resource consumption of the VSCPU with the 80-bit data path (VSCPU-80) design, we designed a VSCPU with the 16-bit data path (VSCPU-16). As shown in Table III, we have made significant changes to the instruction set.

In VSCPU-16 design, an additional register named “W” has been defined to execute the algorithm in a more optimized manner. Since our key value is 80-bit, it is stored in 5 separate data addresses in this 16-bit architecture design. A new instruction called SRW3 was created, which also uses the W register to perform the 3-bit right shift required in the key schedule function. Using this instruction repeatedly, rotation of the key value which is found in different addresses can be made easily.

With the added SBOX16 instruction, the parameter 0 provides the SBOX conversion in the key schedule function only to a 4-bit section, and with the parameter 1, all input values are converted through SBOX operation.

The permutation function is also provided by calling the PER0, PER1, PER2 and PER3 instructions twice in succession. [6] is used to create these instructions.

With the transition from 80-bit architecture to 16-bit, 45 instruction words of 12-bit and 16 data words of 16-bit were needed, and the algorithm computed the result in 1402 cycles. However, the new design can operate with a clock frequency up to 153 MHz and consumes 270 LUTs and 291 FFs. These results have approached pure RTL implementation in terms of FPGA resource consumption, yet it is seen that it consumes 41% more resources than PicoBlaze implementation. However, when the clock frequency and the latency of both

implementations are considered, it is seen that the VSCPU-16's performance is about 211 times better than the PicoBlaze.

F. Other Works

PRESENT is implemented on FPGAs after it's publication by other teams. We examined some of the most recent and promising researches and added their results to Table I and II.

In 2015, Tay *et al.* developed 8-bit hardware architecture of PRESENT algorithm on a Virtex-5 FPGA [11]. They managed to reduce the amount of resources for SBoxes due to 8-bit datapath, and Karnaugh mapping and further factorization of SBoxes. The resulting implementation consumes 222 LUTs and 201 FFs; it can work with a clock frequency of up to 236 MHz; and it can compute a result in 295 cycles.

In 2016, Lara-Nino *et al.* created an architecture based on a 16-bit datapath [12]. By doing that, the implementation of PRESENT on Spartan-6 FPGAs consumes only 226 LUTs and 89 FFs and the design works with a clock frequency up to 172 MHz and with a latency of 132 cycles.

Diehl *et al.* presented the implementation of 6 different ciphers including PRESENT using both custom hardware design and software design with 8-bit microprocessor [13]. There are only 30 native instructions on this soft microprocessor, and the data words and instruction words are 8-bit length. The processor is tailored to the algorithms and unrequired functionality is removed before implementation as in our work. PRESENT implementation on Kintex-7 FPGAs using this custom processor consumes 377 LUTs, achieves a clock frequency up to 542 MHz, and takes 20030 cycles to complete an encryption operation. 156 instruction words and 48 data words are used in the implementation.

Our pure RTL implementation gives better result than [12] in every aspect and consumes almost the same amount of FPGA resources with [11]. However, the throughput of our pure RTL implementation is almost 6 times higher than of the [11]. Moreover, our VSCPU design with 16-bit architecture is more efficient and consumes lower resources than [13].

IV. CONCLUSION

In this work, we aim to make high performance and efficient implementation of any application specific algorithm for devices with low hardware resources in a short time. As an exemplary application, we chose the PRESENT cipher algorithm on IoT. In order to compare the results, this algorithm was implemented on FPGA with pure RTL, Vivado HLS, PicoBlaze, VSCPU and modified VSCPU. In addition, we examined some of the most recent and promising researches. The modified VSCPU implementation, which has been modified by changing the instruction set and architecture,

gave us the best results due to its low resource consumption, high performance and the ability to develop software on it.

In future studies, VSCPU's compiler and simulator can work according to the changing instruction set and architecture and rapid prototyping can be realized completely. In this way, it would be possible to make high performance design in a short time.

ACKNOWLEDGEMENT

This work is partially supported by a TÜBİTAK (The Scientific and Technological Research Council of Turkey) ARDEB 1001 project (no: 117E090). Prof. Sezer Gören is the PI of the project. While Prof. T. Baris Aktemur is the initial co-PI of the project, Prof. H. Fatih Ugurdag is the current co-PI. Abdullah Yildiz and Deniz Iskender are among the research assistants of the project.

REFERENCES

- [1] J. Daemen and V. Rijmen, "AES proposal: Rijndael", 1999.
- [2] "Data Encryption Standard," Federal Information Processing Standards Publication No. 46, National Bureau of Standards, January 15, 1977.
- [3] A. Bogdanov, L.R. Knudsen, G. Leander, C. Paar, A. Poschmann, M.J. B. Robshaw, Y. Seurin, and C. Vikkelsøe, "PRESENT: An Ultra-Lightweight Block Cipher", Cryptographic Hardware and Embedded Systems (CHES), Lecture Notes in Computer Science, pp. 450–466, 2007.
- [4] "Implementation of the PRESENT lightweight block cipher in VHDL", Github 11.08.2019, <https://github.com/huljar/present-vhdl>.
- [5] A. Yildiz, H.F. Ugurdag, B. Aktemur, D. Iskender, and S. Gören, "CPU design simplified," International Conference on Computer Science and Engineering (UBMK), 2018.
- [6] E.B. Kavun and T. Yalcin, "RAM-Based Ultra-Lightweight FPGA Implementation of PRESENT," International Conference on Reconfigurable Computing and FPGAs, 2011.
- [7] "PicoBlaze 8-bit Microcontroller", Xilinx, 11.08.2019, <https://www.xilinx.com/products/intellectual-property/picoblaze.html>.
- [8] D. Bandyopadhyay and J. Sen, "Internet of Things: Applications and Challenges in Technology and Standardization," Wireless Personal Communications, vol. 58, no. 1, pp. 49–69, Sep. 2011.
- [9] "Report on Lightweight Cryptography", National Institute of Standards and Technology Internal Report 8114, March, 2017.
- [10] D. Dinu, A. Biryukov, J. Groszschädel, D. Khovratovich, Y.L. Corre, L. Perrin, "FELICS - Fair Evaluation of Lightweight Cryptographic Systems", 2015.
- [11] J.J. Tay, M.L.D. Wong, M.M. Wong, C. Zhang, and I. Hijazin, "Compact FPGA implementation of PRESENT with Boolean S-Box," in 2015 6th Asia Symposium on Quality Electronic Design (ASQED), Aug 2015, pp. 144–148.
- [12] C.A. Lara-Nino, M. Morales-Sandoval, and A. Diaz-Perez, "Novel FPGA-based low-cost hardware architecture for the PRESENT block cipher", in 2016 Euromicro Conference on Digital System Design, 2016.
- [13] W. Diehl, F. Farahmand, P. Yalla, J. Kaps and K. Gaj, "Comparison of Hardware and Software Implementations of Selected Lightweight Block Ciphers", in 2017 27th International Conference on Field Programmable Logic and Applications, 2017.

Qubit Test Synthesis Processor for SoC Logic

Wajeb Gharibi
University of Missouri
Kansas City, USA
gharibiw@hotmail.com

David Devadze
Batumi Shota Rustaveli State
University,
Batumi, Georgia
david.devadze@bsu.edu.ge

Vladimir Hahanov
Design Automation Department
Kharkov National University of
Radioelectronics
Kharkov, Ukraine
hahanov@icloud.com

Eugenia Litvinova
Design Automation Department
Kharkov National University of
Radioelectronics
Kharkov, Ukraine
litvinova_eugenia@icloud.com

Ivan Hahanov
Design Automation Department
Kharkov National University of Radioelectronics
Kharkov, Ukraine
ivanhahanov@icloud.com

Abstract— A qubit method for synthesizing tests of discrete functions of SoC components is proposed, which leverages Boolean derivatives with respect to a vector description of logic element's behavior in the form of Q-coverage. The primacy of the metrics of mathematical and technological relations in data structure, on which effective algorithms and methods for controlling or data processing are built to achieve the performance of testing processes, is formulated. A vector model or form of Boolean derivatives is introduced, which is used to synthesize deductive matrices in the qubit fault simulation method and to evaluate the quality of test sequences. A tree-driven ATPG processor, represented by a binary tree-graph of xor-elements for parallel processing of parts of the qubit coverage, and data structures of SoC logic for calculating qubit Boolean derivatives are proposed. The proposed data structures and methods are implemented in a software application that focuses on parallel testing the logic functions of digital systems-on-chips using qubit coverage.

Keywords— qubit test synthesis, qubit coverage, deductive fault simulation, Boolean qubit derivative, logic function, SoC components, ATPG processor.

I. STATE OF THE ART

The goal of the research is to reduce the generation time of tests of logical components of a digital system-on-chip through creating algorithms and methods for parallel synthesis of test patterns for single stuck-at faults of digital devices based on parallel processing when taking Boolean derivatives by using qubit coverage for describing logical functions.

Objectives are the following: 1) Create a model of the relationship of order in data structures and computing, which forms the effectiveness of parallel control of large data processing algorithms when solving test problems. 2) Develop a functional model for the qubit description of logic elements of digital systems-on-chips. 3) Synthesize a parallel processor for taking Boolean derivatives based on the use of qubit coverage. 4) Create a qubit method for synthesizing tests based on taking Boolean derivatives with respect to Q-coverage of the logical functions of digital systems-on-chips.

The productivity of computing is determined by the metric of mathematical and technological relations in the data structures, which are the basis of effective algorithms and methods for controlling or data processing for achieving the goals. The data structures of quantum computing differ from classical calculations by the superposition relation between zero and one, the positioning of which is determined at one point of the Hilbert space [1]. Technologically, the superposition is formed by the spins of electrons, which are also positioned at one point of the interatomic space.

It is the superposition of zero and one at a single point in space, which can be extended to a finite number of discrete states, which is the root cause of relations in the organization of data structures for the implementation of parallel methods and algorithms.

As for classical computing, the property of superposition, but not at one point, but dispersed in space, can and should be used to develop efficient parallel algorithms for processing big data represented in unitary codes, which allow superposition of a finite number of discrete states [2].

An important conclusion that follows from the above reasoning is the primacy of relations on data structures for the synthesis of efficient parallel, but secondary, algorithms for controlling or data processing. Any design of a new product should keep in mind the following order relation between categories: 1) Purpose. 2) Relations. 3) Management. 4) Architecture (Infrastructure + Personnel). 5) Observability. 6) Data (Resources). By the way, this relation of order must be strictly followed when creating effective cyber-social computing mechanisms (companies, universities, states), where the "Infrastructure" is complemented by the "Personnel" component.

The interest of scientists and practitioners to new solutions of the problems of testing computing devices is illustrated by the number of publications in the IEEE Xplore library. For example, content search by request (Fault Simulation) gives 37892 works. At the same time, the high-performance deductive analysis method [3] has a total of 51 references to scientific publications.

Nevertheless, the library contains 290 publications focused on Quantum Fault Simulation, including deductive algorithms [4]. Another example is related to the number of works in the field of synthesis and test generation, with 20408 and 64691 publications, respectively. Moreover, the number of qubit or quantum models and methods for generating (1020) and synthesizing (264) tests has increased significantly over the past 5 years [5–8]. Such interest of scientists and industry is associated with improving the quality of computing in all fields of human activity: data centers, social networks, cloud computing, transportation, energy, medicine, construction, astronautics, weapons, process control. With regard to the fault analysis, the state of technical diagnostics can be represented by the relations between data structures, models of fault-free behavior, faults and methods for the synthesis and analysis of tests, shown in Fig. 1. The variety of technologies in the field of testing have been proposed by scientists and practitioners over the past 70 years, which is reflected in tens of thousands of publications, including [9-10]. Recent publications reflect the appearance of parallel models and methods of synthesis and analysis, based on artificial intelligence, cloud services, and quantum data structures [11–13]. Reversible logic circuits are combinations of several special types of quantum k-CNOT gates, for which classical algorithms for synthesizing test patterns are used to detect all single missed faults in a reversible circuit based on k-CNOT gates using the Boolean difference method [14].

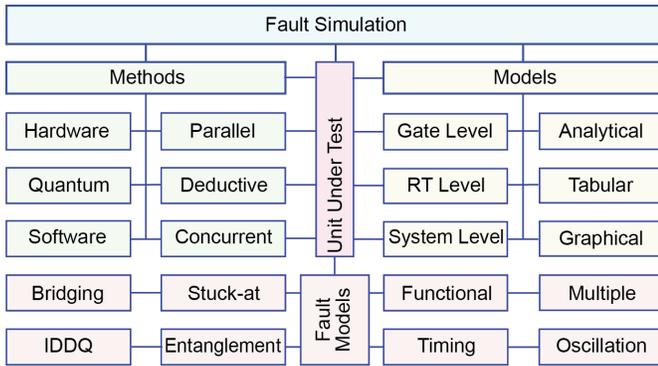


Fig. 1. Models and methods for fault analysis

Test synthesis methods for logical functionalities traditionally use three forms of explicit description of functional behavior: tabular (TT - Truth Table), analytical (DF - Disjunctive Form), graph (alternative Raimund Ubar’s graphs or binary solution diagrams [9]). The first one is technological for computer processing, but time-consuming in terms of memory and table processing performance. The second one is compact in form, but it requires the creation of powerful calculators for analyzing or solving Boolean equations. The third one is visual for humans, compact for our PC, but requires specialized computers for synthesis and analysis of complex systems based on alternative graphs [9].

Next, we propose a qubit coverage [2,4-8], as a vector representation of the state of the outputs of the truth table with an implicit description of the input actions in the form of coordinate addresses of the binary vector of output values. Comparative analysis of the forms for setting the logical

function of three variables, including the qubit vector or Q-coverage (QC - Qubit Coverage), is presented as follows:

$$Y(TT) = \begin{matrix} X_1 & X_2 & X_3 & Y \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{matrix};$$

$$Y(DF) = \bar{X}_1\bar{X}_2X_3 \vee \bar{X}_1X_2\bar{X}_3 \vee X_1\bar{X}_2\bar{X}_3 \vee X_1X_2X_3;$$

$$Y(QC) = 01101001.$$

The Q-vector is a more compact form compared to the truth table, which is characterized by all the advantages of the table related to the technological effectiveness of processing for the synthesis and analysis of logical functions. The Q vector requires less memory in n times to store data compared to the truth table of n variables. The Q-vector does not require (n**2) complicated computational procedures necessary to determine the output state of a logical function using a disjunctive normal form or a generalized truth table. To do this, you need only one automaton operation using addressable write-read operations: $M_i=Q_i[M(X_i)]$, which have parallelism and linear computational complexity.

II. SEQUENTIAL SYNTHESIS OF BOOLEAN DERIVATIVES BY QUBIT COVERAGES

The test synthesis using qubit coverages is based on the technology for determining Boolean derivatives, which create the activation of logical paths from inputs to outputs of a circuit structure. For this, a logical xor-operation between symmetric parts of the qubit vector is used [2]:

$$Q'(X_k) = \{Q_i^L, Q_i^R\} = Q_i^L \bigoplus_{i=1,2}^{k-1} Q_i^R.$$

The equation can be put in correspondence with the simplest sequencer for taking the qubit derivative (Fig. 2), where the xor-operation forms both parts of the resulting vector.

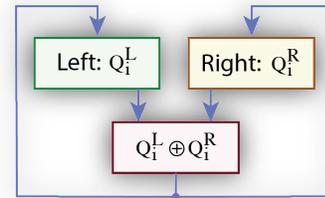


Fig. 2. Sequencer of qubit derivative

Using a sequencer to take three qubit derivatives for a logic function of three variables $Q = (01\ 10\ 10\ 01)$ is presented as follows:

$$Q(X1) = (11\ 11\ 11\ 11); \quad Q(X2) = (11\ 11\ 11\ 11);$$

$$Q(X3) = (11\ 11\ 11\ 11).$$

The first vector of the qubit derivative with respect to X1 is formed as the xor-sum of adjacent cells, the second vector of the qubit derivative with respect to X2 is formed as the xor-sum of adjacent coordinate pairs, the third vector of the qubit derivative with respect to the X3 is synthesized as the xor-sum of the adjacent quadruple of the qubit vector. The consequences for such X-functions are interesting. If the qubit derivatives with respect to all variables of the logical function are equal to the unit vector, then: 1) The deductive formula for fault simulation is invariant to the input test sets or does not depend on them. 2) Any test vector detects all single stuck-at faults inverse to the line states. 3) The number of logical functions of n variables where the following condition is true

$$\prod_{i=1}^n \left[\frac{df}{dX_i} = 1 \right],$$

always equal to two. 4) To activate the input variable Xi, no logical conditions are needed for other inputs. 5) The X-function is a simple logical function of a finite number of variables (n = 1,2,3, ...), which cannot be minimized. 6) The length of the test for detecting single stuck-at faults of all lines for the X-function of n variables is always equal to

$$Q = 1 + \frac{1}{2} \times 2^{2^n}.$$

The following tables [6] represent the process of test synthesis (T) using a qubit coverage Q = (01 10 10 01), fault simulation (D), minimization of test patterns (M) and obtaining a minimum qubit test of a digital structure - table T(Q):

T	1	2	3	4	5	6	7	8
0	0	0	0	0	0	0	0	0
1	0	0	1	1	0	0	0	1
2	0	1	0	0	1	0	0	1
3	0	1	1	0	0	0	0	0
4	1	0	0	0	1	0	1	1
5	1	0	1	0	0	0	0	0
6	1	1	0	0	0	0	0	0
7	1	1	1	0	0	0	1	1

D	1	2	3	4	5	6	7	8
0	1	1	1	1	1	1	1	1
1	1	1	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
3	1	0	0	1	1	1	1	1
4	0	1	1	0	0	0	0	0
5	0	1	0	1	1	1	1	1
6	0	0	1	1	1	1	1	1
7	0	0	0	0	0	0	0	0

M	1	2	3	4	5	6	7	8
0	1	1	1	1	1	1	1	1
1	1	1	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
4	0	1	1	0	0	0	0	0
7	0	0	0	0	0	0	0	0
C	x	x	x	x	x	x	x	x

T(Q)
1
1
1
0
1
0
0
1

The column T(Q) = 11101001 defines the minimum qubit test form – binary addresses for unity coordinates, which need to be entered on the external inputs in order to detect all the single stuck-at faults of the external and internal lines of the logic circuit.

III. QUBIT DERIVATIVE PROCESSING FOR SOC LOGIC TEST GENERATION

Initially, the characteristic equation for the synthesis of tests uses the following operations: negation, towards-shift of data, xor-addition and disjunction on the qubit coverage [2,6]:

$$T(S) = \prod_{j=1}^n [Q \oplus S_j(\bar{Q})].$$

The computational complexity of the test synthesis algorithm for this equation is equal to

$$Q = 2^n + n \times 2^n + n \times 2^n + n \times 2^n = 2^n + 3(n \times 2^n) = 2^n(1 + 3n).$$

The apparatus of qubit Boolean derivatives greatly simplifies the algorithm for generating test patterns for logical functions up to two operations (taking the derivative with respect to each variable and the disjunction of the derivatives, which creates the qubit form of the test):

$$T(Q') = \prod_{i=1}^n Q'(X_i)$$

The Boolean derivative on the qubit vector is reduced to performing xor-operations on parts of the qubit coverage, the dimension of which is determined by the power of two from the number of the variable under consideration, which varies from 1 up to n. The following qubit coverage describes the behavior of a logic function of four variables: Q(X) = (1000000000000001), for which taking derivatives for test synthesis is considered.

The characteristic equation of taking the Boolean derivative operates with parts of the qubit coverage vector, the dimension of which is associated with the power of two on the number of variables. Next, we consider an algorithm for taking the qubit derivative of a function of 4 variables:

1) At the first step, the derivative with respect to the variable X1 is taken: two equal parts of the binary Q-vector of the dimension 2^n are considered, to which the parallel (coordinate-wise) xor-operation is applied: Q0 ⊕ Q1 → {Q0, Q1}, after which the result is entered in both parts of the Q-vector. Otherwise, this procedure can be written as Q0 = Q0 ⊕ Q1, Q1 = Q0 ⊕ Q1 or in compact form {Q0, Q1} = Q0 ⊕ Q1.

2) At the second step, the derivative with respect to the variable X2 is taken: four equal parts of the qubit Q-vector are already considered, to which two xor-operations are applied in pairs, the result of which is entered into operands: {Q0, Q1} = Q0 ⊕ Q1; {Q2, Q3} = Q2 ⊕ Q3.

3) At the third step, the derivative with respect to the variable X3 is taken: eight equal parts of the qubit Q-vector are already considered, to which four xor-operations are applied in pairs, the result of which is entered into operands:

$$\{Q_0, Q_1\} = Q_0 \oplus Q_1; \{Q_2, Q_3\} = Q_2 \oplus Q_3;$$

$$\{Q_4, Q_5\} = Q_4 \oplus Q_5; \{Q_6, Q_7\} = Q_6 \oplus Q_7.$$

4) At the fourth step, the derivative with respect to the variable X4 is taken: 16 equal parts of the qubit Q-vector are already considered, to which eight xor-operations are applied in pairs, the result of which is entered into operands:

$$\{Q_0, Q_1\} = Q_0 \oplus Q_1; \{Q_2, Q_3\} = Q_2 \oplus Q_3;$$

$$\{Q_4, Q_5\} = Q_4 \oplus Q_5; \{Q_6, Q_7\} = Q_6 \oplus Q_7;$$

$$\{Q_8, Q_9\} = Q_8 \oplus Q_9; \{Q_{10}, Q_{11}\} = Q_{10} \oplus Q_{11};$$

$$\{Q_{12}, Q_{13}\} = Q_{12} \oplus Q_{13}; \{Q_{14}, Q_{15}\} = Q_{14} \oplus Q_{15}.$$

The structure of the processor for taking Boolean derivatives by qubit coverage is shown in Fig. 3 as a binary tree-graph of function's Q-vector representation, where the nodes are xor-converters, and the edges are the operands or parts of the qubit coverage.

The right part of the Fig. 3 represents the data structures transformed using processor xor-operations. The feature of the processor is the parallel execution of all operations on the data represented by qubit coverage of logical functionality. Therefore, the xor-elements of the graph and the corresponding registers of the data structures have end-to-end connection to each level from the qubit coverage of a logic function located in the bottom row of both structures. There are interesting parallel solutions in the analysis of digital devices for test synthesis [15-17]. However, there are no analogs of computational architectures for the performance when taking Boolean derivatives in one automaton cycle.

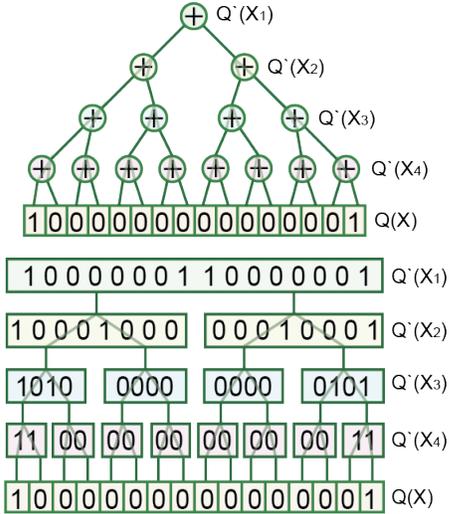


Fig. 3. Tree-driven ATPG processor and data structures for SoC logic

Table 1 illustrates the taking of four Boolean derivatives based on the use of the proposed processor, which are further combined into a qubit test.

TABLE 1. BOOLEAN DERIVATIVES, TEST PATTERNS AND TRUTH TABLE

Derivative-based Test Pattern Generation	
Q(X)	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
Q'(X ₁)	1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1
Q'(X ₂)	1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1
Q'(X ₃)	1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1
Q'(X ₄)	1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1
$T(Q) = \bigvee_{i=1}^n Q'(X_i)$	1 1 1 0 1 0 0 1 1 0 0 1 0 1 1 1 1

Truth Table	
Q(X)	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
X ₁	0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1
X ₂	0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 1
X ₃	0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 1
X ₄	0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1

The truth table is given for the convenience of representation of the analytical form of defining the derivatives with respect to the states of variables specified in the truth table. Considering the above, the analytic disjunctive normal form of derivatives with respect to four variables:

$$Q'(X_1) = (1000000110000001);$$

$$Q'(X_2) = (1000100000010001);$$

$$Q'(X_3) = (1010000000000101);$$

$$Q'(X_4) = (1100000000000011)$$

is transformed to the following:

$$Q'(X_4) = \bar{X}_1\bar{X}_2\bar{X}_3\bar{X}_4 \vee \bar{X}_1\bar{X}_2\bar{X}_3X_4 \vee X_1X_2X_3\bar{X}_4 \vee X_1X_2X_3X_4 = \bar{X}_1\bar{X}_2\bar{X}_3 \vee X_1X_2X_3.$$

$$Q'(X_3) = \bar{X}_1\bar{X}_2\bar{X}_3\bar{X}_4 \vee \bar{X}_1\bar{X}_2\bar{X}_3X_4 \vee X_1X_2\bar{X}_3X_4 \vee X_1X_2X_3X_4 = \bar{X}_1\bar{X}_2\bar{X}_4 \vee X_1X_2X_4.$$

$$Q'(X_2) = \bar{X}_1\bar{X}_2\bar{X}_3\bar{X}_4 \vee \bar{X}_1\bar{X}_2\bar{X}_3X_4 \vee X_1\bar{X}_2X_3X_4 \vee X_1X_2X_3X_4 = \bar{X}_1\bar{X}_3\bar{X}_4 \vee X_1X_3X_4.$$

$$Q'(X_1) = \bar{X}_1\bar{X}_2\bar{X}_3\bar{X}_4 \vee \bar{X}_1\bar{X}_2X_3X_4 \vee X_1\bar{X}_2\bar{X}_3\bar{X}_4 \vee X_1X_2X_3X_4 = \bar{X}_2\bar{X}_3\bar{X}_4 \vee X_2X_3X_4.$$

The analytical form of the result has no lines, with respect to which the derivative is taken, since the equations show binary conditions for activating variables on combinations of other lines, forming two logical paths from each input to the output, which is the basis for the test synthesis for stuck-at faults. Summarizing the above, the change of each variable of Q-derivatives ensures the detection of all single stuck-at faults on the input, internal and output lines on all possible logical activation paths. Having a qubit coverage of an arbitrarily complex Boolean function, the qubit test synthesis method is reduced to one parallel xor-operation for the synthesis of derivatives on parts of a Q-vector, which makes it possible to get a test for single stuck-at faults in n-1 automaton cycle by combining the vectors of the obtained derivatives.

The computational complexity of the basic test synthesis algorithm based on taking qubit derivatives is $Q = n \times 2^n + n \times 2^n = Q = 2(n \times 2^n)$. When using a processor, the complexity of the test synthesis algorithm is reduced to n automaton clock cycles $Q = 1 + (n - 1) = n$. However, the Q-test obtained on the basis of uniting qubit derivatives does not guarantee its minimality.

As for test synthesis for X-functions (xor, not-xor) [2,4], there is no need to perform calculations at all. You only need to generate a test of the length

$$Q = 1 + \frac{1}{2} \times 2^{2^n}.$$

using Q-coverage in accordance with the following rules – the qubit test is equal to the qubit coverage of the X-function, where any 0-coordinate is additionally replaced with a unit value:

$$T = T^1 \vee T_i^0,$$

$$T^1 = \bigvee T_i : f(T_i) = 1$$

$$T_i^0 \in T^0 = \bigvee T_i : f(T_i) = 0;$$

$$T(01101001) = (001 \vee 010 \vee 100 \vee 111) \vee 000;$$

$$T(10010110) = (000 \vee 011 \vee 101 \vee 110) \vee 001.$$

Thus, tests for qubit coverages of two Boolean X-functions $Q1 = 01101001$ and $Q2 = 10010110$ are the vectors: $T1 = 11101001$ and $T2 = 11010110$.

IV. CONCLUSION

1) A model of order relations in data structures and computing is created, which forms the efficiency of parallel control of large data processing algorithms through the superposition of a finite set of discrete states.

2) A structural model of the interaction of qubit coverages of logical functions and derivatives of the components focused on the synthesis and analysis of digital systems in order to obtain test patterns for single stuck-at faults has been developed in order to reduce the design and testing time of computing devices.

3) The concept of simple X-functions of a finite number of variables is introduced, which are characterized by the absence of minimization and the presence of testability properties, which makes it possible to synthesize digital devices, technologically advanced for solving test, simulation and diagnosis problems.

4) A tree-driven ATPG processor and data structures for SoC logic are proposed for calculating qubit Boolean derivatives, represented by a binary tree-graph of xor-elements, for parallel processing of parts of the qubit coverage.

5) Qubit methods for test synthesis of logical functions are proposed, which are characterized by the quadratic and linear computational complexity of algorithms for generating test sequences.

6) Further research is related to the creation of technological solutions for parallel test synthesis and development of a processor for the synthesis of matrices of deductive parallel analysis of test pattern quality by using qubit coverages of logic circuits.

REFERENCES

[1] M.A. Nielsen, I.L. Chuang, "Quantum Computation and Quantum Information," Cambridge University Press, 2010.

[2] V. Hahanov, "Cyber Physical Computing for IoT-driven Services," Springer, New York, 2018.

[3] U. Reinsalu, J. Raik and R. Ubar, "Register-transfer level deductive fault simulation using decision diagrams," 2010 12th Biennial Baltic Electronics Conference, Tallinn, 2010, pp. 193-196.

[4] V. Hahanov, A. V. Hacimahmud, E. Litvinova, S. Chumachenko and I. Hahanova, "Quantum Deductive Simulation for Logic Functions," 2018

IEEE East-West Design & Test Symposium (EWDTS), Kazan, 2018, pp. 1-7.

[5] V. Hahanov, W. Gharibi, S. Chumachenko, E. Litvinova, I. Iemelianov, M. Liubarskyi, "Quantum Data Structures for SoC Component Testing," International Journal of Design, Analysis & Tools for Integrated Circuits & Systems, Oct. 2017, vol. 6, iss. 1, P. 23.

[6] V. Hahanov, E. Litvinova, S. Chumachenko, I. Iemelianov, M. Liubarskyi, "Qubit test synthesis for the black box functionalities," Proc. of 5th Prague Embedded Systems Workshop, June 29-30, 2017, Roztoky u Prahy, Czech Republic, pp.45-51.

[7] V. Hahanov, W. Gharibi, M. Liubarskyi, E. Litvinova, M. Gharibi, S. Chumachenko, I. Hahanov, and A. Hahanova, "Quantum Deductive Fault Simulation," Proc. of Int'l Conf. Modeling, Sim. and Vis. Methods (MSV'18) in The 2018 World Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE2018), Jul 30-Aug 02, 2018, Las Vegas, USA, pp. 10-16.

[8] V. Hahanov, S. Chumachenko, I. Hahanova, I. Iemelianov, I. Hahanov, "Quantum Sequencer for the Minimal Test Synthesis of Black-box Functionality," Proc. of IEEE East-West Design and Test Symposium, October, 2017, Novi Sad, pp. 445-450.

[9] J. Raik and R. Ubar, "Sequential circuit test generation using decision diagram models," Design, Automation and Test in Europe Conference and Exhibition, 1999. Proceedings (Cat. No. PR00078), Munich, Germany, 1999, pp. 736-740.

[10] A. Shrestha, L. Xing and Y. Dai, "Decision Diagram Based Methods and Complexity Analysis for Multi-State Systems," in IEEE Transactions on Reliability, vol. 59, no. 1, pp. 145-161, March 2010.

[11] V. Dhare and U. Mehta, "Defect characterization and testing of QCA devices and circuits: A survey," 2015 19th International Symposium on VLSI Design and Test, Ahmedabad, 2015, pp. 1-2.

[12] A. Abdollahi and M. Pedram, "Analysis and Synthesis of Quantum Circuits by Using Quantum Decision Diagrams," Proceedings of the Design Automation & Test in Europe Conf., Munich, 2006, pp. 1-6.

[13] V. Dhare and U. Mehta, "A Simple Synthesis Process for Combinational QCA Circuits: Q-Synthesizer," 2019 32nd International Conference on VLSI Design and 2019 18th International Conference on Embedded Systems (VLSID), Delhi, NCR, India, 2019, pp. 498-499.

[14] B. Mondal, D. K. Kole, D. K. Das and H. Rahaman, "Generator for Test Set Construction of SMGF in Reversible Circuit by Boolean Difference Method," 2014 IEEE 23rd Asian Test Symposium, Hangzhou, 2014, pp. 68-73.

[15] M.G. Whitney, "Practical Fault Tolerance for Quantum Circuits," A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Computer Science in the Graduate Division of the University of California, Berkeley, 2009.

[16] J.P. Hayes, I. Polian, B. Becker, "Testing for Missing-Gate Faults in Reversible Circuits," Proc. Asian Test Symposium, Taiwan, November 2004.

[17] R. J. Lipton, K. W. Regan, "Quantum Algorithms via Linear Algebra," MIT Press eBook, 2014.

Increasing the Effective Volume of Digital Watermark Used in Monitoring the Program Code Integrity of FPGA-Based Systems

Kostiantyn Zashcholkin
*Department of Computer Intelligent
Systems and Networks
Odessa National Polytechnic
University*
Odessa, Ukraine
const-z@te.net.ua

Oleksandr Drozd
*Department of Computer Intelligent
Systems and Networks
Odessa National Polytechnic
University*
Odessa, Ukraine
drozd@ukr.net

Ruslan Shaporin
*Department of Computer Intelligent
Systems and Networks
Odessa National Polytechnic
University*
Odessa, Ukraine
rshaporin@gmail.com

Olena Ivanova
*Department of Computer Systems
Odessa National Polytechnic University*
Odessa, Ukraine
en.ivanova.ua@gmail.com

Yulian Sulima
*Computer Systems Department
Odessa Technical College of the Odessa
National Academy of Food Technologies*
Odessa, Ukraine
mr_lemur@ukr.net

Abstract—The paper deals with monitoring the program code integrity of FPGA-based systems. An approach to the integrity monitoring considered in the paper is based on embedding the digital watermark into the information object of FPGA chips program code. A digital watermark, which is embedded into the program code, contains a monitoring hash sum. Such kind of embedding does not change the program code size and operation of device, which is programmed with the help of this program code. In using this approach the integrity monitoring is provided by the condition that the digital watermark extraction and the recovery of initial state of program code information object occur simultaneously. In the paper a method, which allows increasing the embedded digital watermark effective volume, is proposed. Increasing the effective volume of the digital watermark gives the possibility to use a broader set of hash functions to monitoring the integrity. This allows using the hash functions possessing a big cryptographically strong in the process of the integrity monitoring. Increasing the effective volume is achieved due to the preliminary preparation of the FPGA program code information object, which (preparation) is performed before the embedding of digital watermark. In the course of this preparation the information object bits set by the embedding key are led to some predetermined state. In the paper an experimental research of the proposed method efficiency in the point of increasing the effective volume of the digital watermark is presented.

Keywords—*Integrity Monitoring, Integrity Analysis, Digital Watermarks, FPGA-Based Systems, LUT-Oriented Architecture, Program Code of FPGA*

I. INTRODUCTION

At the moment a considerable share of hardware of the computer and control digital systems is based on programmable devices. One of the main reasons of preference of the very programmable devices is that there is the possibility to modify their operation during all the life cycle. Due to this possibility a number of typical tasks, which can appear at the different stages of life cycle of the computer or control system, is solved simply enough (as compared to nonprogrammable devices). We can refer to such kinds of tasks the following ones: a) functioning faults elimination detected in the course of the device operation;

b) expansion and changing the set of functions, which are provided by the device; c) functioning optimization of the device.

However the possibility to modify the program code of programmable devices generates the problem of provision of this program code integrity [1]. The potential accessibility to the program code rewriting function is the basis for vulnerability, which allows to illegitimately bring modification to the program code [2]. The presence of legitimate program code modification in the process of the device operation permits to mask a malicious modification presenting it as a part of a legitimate one [3]. The program code integrity violation of devices, entering the composition of systems of both safety-critical [4] and mass usages [5], creates the excessive risk with unacceptable consequences [6, 7]. So the safety of systems, in which the programmable devices are included, cannot be ensured without the solution of problem of the program code integrity provision.

In the given paper a problem of the program code integrity provision of one of the widely used classes of programmable devices – FPGA chips (Field Programmable Gate Array) [8] is considered. The FPGA chips are a set of programmable basic calculating units, the links between which are ensured by the programmable system of commutation. The natural parallelism of the computing tasks solution with the help of FPGA chips creates their (chips) advantage [9, 10] in performance characteristics as compared to microprocessors.

In spite of the presence of embedded mechanisms of the program code protection from rewriting in many FPGA-based systems there are the bypass ways of such kind of protection allowing to enter the illegitimate modification in the program code [11, 12]. By virtue of this the most popular approaches to the provision of the program code integrity of FPGA-based components is a combination of processes of access restriction to the program code and integrity monitoring. The integrity monitoring is traditionally based on the usage of extra monitoring data units allowing to make the conclusions about the code integrity.

II. LITERATURE REVIEW AND GOAL OF THE PAPER

The most popular approaches to the program code integrity monitoring used in practices have become the ones, which use a hash sum [13]. For the program code information object a hash sum is calculated with the help of the set hash function [14]. This hash sum is considered further to be a standard one. A standard hash sum is in some way matched with the program code information object or joins it. Further if checking the program code integrity is to be executed the recalculation of information object hash sum is implemented. The comparison of the standard and newly calculated hash sum permits to confirm the integrity or detect its violation.

One of the substantial constituents of the integrity monitoring efficiency (in the point of counteraction to the attempts to bypass monitoring) is a way and location of the standard hash sum storage. In using the traditional approaches to the integrity monitoring the following ways (or their variations) of storing the standard hash sum are applied.

1) A standard hash sum is stored separately [15] from the program code information object in some centralized database. The main disadvantage of this way of storing is the complexity of the database protection from information leakage. The mass leakage of information from database with hash sum (which is a quite frequent event as the practice shows) compromises all the systems of integrity monitoring, which this database provides [16]. Even under the conditions of extra encryption of the standard hash sum the access to its encrypted values creates a potential chance of spoofing and bypassing the integrity monitoring [17].

2) The standard hash sum is stored together [18] with the program code information object in the FPGA configuration memory. The disadvantage of this way is conditioned, firstly, with the evidence for an outside surveillance that the integrity monitoring of the given information object is carried out, and secondly, that the standard hash sum is accessible and this makes the attempts to spoof it easier.

3) The standard hash sum is included [19] in the program code information object and stored as its constituent. The hash sum detection inside the information object is not of great difficulty because the hash sum is not distributed about

the information object but is centrally stored in its structure. By virtue of this the given way has the disadvantages similar to the previous one.

Thus the described ways of storing a standard hash sum potentially create the vulnerability, which can become a cause to attempt to spoof the hash sum with the aim to hide the integrity violation.

A perspective approach to integrity monitoring is the standard hash sum embedding into the program code information object in the form of a digital watermark [20, 21]. Such kind of approach masks from an outside surveillance the very fact of the integrity monitoring implementation [22]. The digital watermark imbedding does not change the size of the program code information object [23]. Moreover as a result of the digital watermark embedding the operation of programmable device, which functioning is set by the program code, is not modified. These features of the digital watermark are the results of usage of the special equivalent conversion with respect to program code elements. For embedding the digital watermark into FPGA an equivalent conversion of program code of the series-connected LUT (Look Up Table) [24, 25] basic calculating units is used [26, 27]. Wherein the system of links between LUT units as well as the operation, energy characteristics and performance of the device are not changed [28, 29].

The digital watermark extraction from the FPGA program code is possible if steganographic key [30] is available. The key determines the rules of the digital watermark bits placement in the LUT unit set.

The peculiarity of integral monitoring, which is carried out with the help of the digital watermark, is the necessity to recover an initial state of the program code information object [31, 32]. At the moment of monitoring execution (Fig. 1) the digital watermark (containing hash sum) is to be extracted from information object, and the information object itself is to be recovered in the state, in which it existed prior to embedding the digital watermark (initial state). Such recovery is necessary because the standard hash sum is calculated for the initial state of information object. Embedding the digital watermark changes this state.

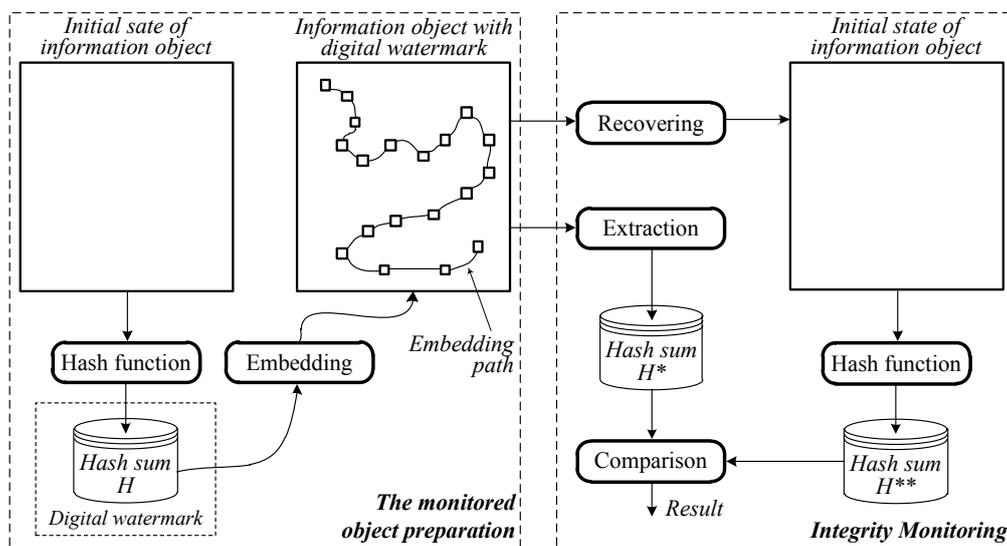


Fig. 1. Scheme of the integrity monitoring, which is based on the digital watermark usage

To ensure the initial state recovery of the program code information object a compression-based approach is used. In Fig. 2 the basic principle of this approach is shown.

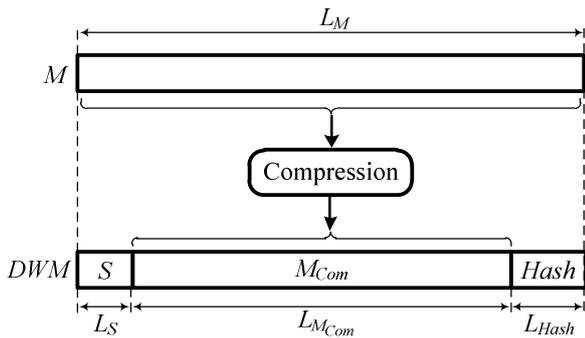


Fig. 2. Compression-based approach to save the initial state of the program code information object

The bit values $M = \langle m_1, m_2, \dots, m_n \rangle$ (values of the specify bits of the LUT units program code) of information object, which are along the embedding path of digital watermark, are combined in a bit sequence. This bit sequence is subjected to the lossless compression procedure [33, 34]. The compressed bit sequence M_{Com} together with service data S (which contains the fields length of the digital watermark) and the standard hash sum creates the digital watermark DWM . This digital watermark is embedded into the place of the bit sequence M by the equivalent conversion [23, 24]. Thus the standard hash sum size in the digital watermark cannot exceed value (1).

$$L_{Hash} = L_M - L_{M_{Com}} - L_S \quad (1)$$

The size of service data S field is fixed. On this basis the size L_{Hash} is dependent on: the number of LUT units (the sequence M length), in which the digital watermark embedding is executed; the applied compression method; the content of the bit sequence M . Wherein the value L_{Hash} can be learnt after indicating the location of watermark embedding and bit sequence M compression.

In case if the amount of bits necessary for storing the standard hash sum exceeds the value L_{Hash} a situation arises when the information object cannot be finally prepared for integrity monitoring. In this case a hash function, which gives a hash sum with less number of bits, should be chosen. If such kind of hash-function change is inaccessible in accordance with the monitoring conditions one should give up monitoring with the help of the digital watermark. On this basis we can constant the following issues: a) field size limitation of the monitoring digital watermark, which is dedicated for storing the standard hash sum; b) instability of this size and possibility to learn it only at the final stages of the information object preparation for integrity monitoring.

The goal of the given paper is to increase the effective volume (which is intended for storing the monitoring hash sum) of the digital watermark as compared to the integrity monitoring methods using compression for the recovery of information object state.

III. THE INTEGRITY MONITORING METHOD PROVIDING THE INCREASED EFFECTIVE VOLUME OF THE DIGITAL WATERMARK

We offer a method of the FPGA program code integrity monitoring, which allows like all compression-based methods:

- to save the initial state of the FPGA program code information object (at the stage of preparing the information object for monitoring);
- to execute the initial state recovery of the FPGA program code information object and the digital watermark extraction simultaneously (at the stage of integrity monitoring implementation).

However wherein the proposed method allows providing the larger effective volume of the digital watermark (the volume intended for the hash sum monitoring storage) than the one (volume) provided by compression-based methods.

That is why the initial state recovery by means of the preliminary preparation of the FPGA program code information object is offered. The proposed method uses some Wong's method [35, 36] ideas as a base. According to the method offered by Wong a fragile digital watermark is embedded into a bitmap image. The property of the digital watermark fragility in the method by Wong makes possible the image integrity monitoring. This method also requires to bring some bits in the values of image pixels to the predetermined state.

The proposed method in the given paper differs from the one by Wong in the following aspects.

The proposed method is oriented to the digital watermark embedding into the FPGA program code and permits only the equivalent conversion of basic units values of the information object. But the method by Wong is oriented to the digital watermark embedding into a multimedia information object and allows the distortion of the basic unit values of this object.

The method by Wong requires to mandatorily bring all basic unit values of the information object to predetermined state. The method offered in the given paper requires to bring only the basic unit values, which are along the embedding path of the digital watermark, to the predetermined state.

The method by Wong fixes the least significant bits as target embedding bits (this is conditioned by the peculiarity of multimedia information objects the method by Wong is oriented to). The proposed method gives the possibility to use equally any of the bits of the basic units (LUT units) program code of information object.

The method by Wong indicates only a single rule how to bring the target bits to the predetermined state – their setting in value 0. The proposed method allows to use any determinate rules (described in the corresponding steganographic key component) to bring the target bits to the predetermined state.

To formulate the principles of the proposed method the following notations and definitions are introduced.

Let $L = \{LUT_1, LUT_2, \dots, LUT_p\}$ is a set of LUT units of FPGA-based device, in the program code of which the monitoring digital watermark is embedded.

On the basis of the rules indicated by steganographic key an ordered set of LUT units, which are along the embedding path of the digital watermark $EmbPath = \langle l_1, l_2, \dots, l_n \rangle$, is formed from this set. In the course of embedding the digital watermark bits are directly embedded into the program codes of the $EmbPath$ LUT units.

Each of the units $l_i \in EmbPath$, where $i=1 \dots n$ contains k -bit program code P_i , respectively. In each of the program codes P_i one of the bits d_i of monitoring digital watermark can be embedded with the help of equivalent conversion [23, 24].

To each of the units $l_i \in EmbPath$ (which is along the embedding path) corresponds one bit $m_i \in P_i$. This bit of program code P_i can be used for embedding one bit of the digital watermark. The correspondence between $l_i \in EmbPath$ and $m_i \in P_i$ is set by rules described in steganographic key. Below the bits m_i will be called the *target bits of embedding*.

The basic theoretical principles of the proposed method are as follows.

The first principle of the method: the initial state recovery of FPGA program code information object is provided on account of the preliminary preparation of this information object. The preparation is carried out prior to embedding the digital watermark into information object. The preparation lies in bringing the target bits of embedding to some predetermined state. The state, which these bit values are brought to, is indicated by rules including in the steganographic key structure.

The second principle of the proposed method: bringing the target bits to the predetermined state (set by steganographic key) is performed with the help of the equivalent conversion [23, 24] similar to those, which are used for the digital watermark embedding.

The third principle of the proposed method: the digital watermark within the framework of the proposed method contains only the monitoring hash sum. There is no information for initial state recovery of information object in it. The lack of necessity to save this information is conditioned by the fact that the initial state of information object is recovered according to the rules described in steganographic key.

The fourth principle of the proposed method: steganographic key (which is used in embedding and extracting the digital watermark) contains the rules for bringing the target bits to the predetermined state. These rules regulate both values themselves (fixed or changed according to some law) and their location in the space of FPGA program codes of LUT units.

To provide this principle a component, which describes the rules of bringing the target bits to the predetermined state, is offered to include in steganographic key:

$$PD\text{-rule} = \langle \text{value}, \text{location} \rangle,$$

where $\text{value} \in \{\text{fixed-value}, \text{value-pattern}, \text{random-value-rule}\}$; $\text{location} \in \{\text{fixed-location}, \text{location-pattern}, \text{iteration-location-rule}, \text{random-location-rule}\}$.

Component *PD-rule* consists of two elements: element *value* indicates a rule of the target bits value formation in the course of their bringing to the predetermined state; element *location* sets the target bits location in the space of LUT units program codes.

Element *value* determines three possible ways of the target bits value formation: *fixed-value* is fixed value 0 or 1 for all the target bits; *value-pattern* is the values, the changes of which are described by some regular pattern; *random-*

value-rule is the values, the changes of which are set by a rule based on pseudo-random number sequence.

Element *location* determines four possible ways of specifying the target bits location: *fixed-location* is the location in bits, which have one and the same number in all LUT units program codes; the rest three ways set the location in bits, number of which changes from unit to unit in accordance with some rule; *location-pattern* is a regular pattern of the bit number change; *iteration-location-rule* is an iteration rule of the bit number change; *random-location-rule* is a rule of the bit number change based on the pseudo-random sequence.

The proposed method is a sequence of stages which are performed in preparing the FPGA program code information object for integrity monitoring, as well as the ones, which are executed in the course of the monitoring itself.

The preparations of FPGA code information object for integrity monitoring.

Stage 1. According to the rules included in the steganographic key components the units, which are along the embedding path, are chosen from the set of LUT units. These units create the ordered sequence $EmbPath = \langle l_1, l_2, \dots, l_n \rangle$.

Stage 2. In accordance with the steganographic key component $location \in PD\text{-rule}$ the ordered sequence of target bits $M = \langle m_1, m_2, \dots, m_n \rangle$ is formed wherein each target m_i is a bit of the LUT unit l_i program code.

Stage 3. In accordance with the steganographic key component $value \in PD\text{-rule}$ the binary values sequence $A = \langle a_1, a_2, \dots, a_n \rangle$ is formed. These values are considered to be the initial ones for target bits of the digital watermark embedding.

Stage 4. With the help of the equivalent conversions [23, 24] the target bits $m_i \in M$ replacement with the initial values $a_i \in A$ is performed. After this FPGA program code information object is considered to be brought to the initial predetermined state.

Stage 5. For the program code information object a monitoring hash sum is calculated. This hash sum is calculated with the help of a hash function, set by steganographic key.

Stage 6. The obtained hash sum is embedded into the target bits of the FPGA program code information object in the form of digital watermark. The embedding is performed according to the traditional methods of the digital watermark embedding into FPGA program code [26, 30].

The executions of information object integrity monitoring.

Stage 1. In accordance with the rules determined by the steganographic key components the units, which are along the embedding path, are chosen from the LUT units set.

Stage 2. According to the steganographic key component $location \in PD\text{-rule}$ an ordered sequence of target bits is formed. At the stage of information object preparation the digital watermark is embedded into these bits.

Stage 3. The digital watermark, which contains the monitoring hash sum, is extracted from these bits.

Stage 4. An action analogous to the one, which is performed at *Stage 3* in preparing the information object for monitoring, is carried out: in accordance with the steganographic key component $value \in PD\text{-rule}$ the binary values sequence $A = \langle a_1, a_2, \dots, a_n \rangle$ is formed.

Stage 5. The initial state recovery of the program code information object is performed. To do this the target bits $m_i \in M$ replacement with the initial values $a_i \in A$ is implemented with the help of the equivalent conversion [23, 24].

Stage 6. For the program code information object obtained at *Stage 5* the monitoring hash sum is calculated. This hash sum is calculated with the help of a hash function, set by steganographic key.

Stage 7. The comparison of hash sum extracted from the information object at *Stage 3* and the one calculated at *Stage 6* is performed. If these hash sums coincide the information object integrity is considered to be confirmed. Otherwise the integrity violation is fixed.

IV. THE PROPOSED METHOD AND EXPERIMENT DISCUSSION

The proposed method efficiency in the point of effective volume increase of the digital watermark is as follows. The traditional methods of integrity monitoring (which are based on the digital watermark usage) apply the compression to save the initial information object state. These methods permit to use only a small part (1) of the digital watermark

volume for storing the monitoring hash sum. This reason does not allow for in some cases the traditional methods to provide the saving of hash sum with a size necessary for monitoring. As to the method proposed in the presented paper it gives the possibility to use all the available volume of the digital watermark for storing the hash sum.

To compare the offered method to the traditional ones an experiment was made. The experiment was made for five FPGA projects of different volume. The synthesis of these projects was implemented in CAD environment Intel (Altera) Quartus [37] for target FPGA chips Intel Cyclone IV [38, 39].

For all the five projects the embedding path formation was performed with the help of one and the same steganographic key. Then the authors indicated what size of the hash sum is which can be provided by a traditional compression-based method of integrity monitoring. The sequence of target bits was formed with further performing the compression of this sequence. Then according to equation (1) the maximal possible size of hash sum was calculated. We also indicated which of the most popular [40, 41] hash functions can provide a hash sum that could fit to this possible size.

The results are presented in table 1 (the projects are ordered according to the total amount of LUT units).

TABLE I. EXPERIMENT RESULTS

Project No	Total amount of LUT units in project	Amount of LUT units, which are along the embedding path	Traditional compression-based method		Proposed method: possibility to use hash functions (size)
			Maximum possible amount of hash sum bits	Possibility to use hash functions (size)	
1	780	143	6	—	MD5 (128)
2	4212	904	35	—	SHA1(160) or MD5 (128)
3	10074	2851	133	MD5 (128)	SHA1(160) or MD5 (128)
4	11839	3179	141	MD5 (128)	SHA1(160) or MD5 (128)
5	15043	4811	168	SHA1(160) or MD5 (128)	SHA1(160) or MD5 (128)

From table 1 one can see that the traditional method does not give the possibility to save a suitable size hash sum (obtained with the help of some high-usage hash function) for projects 1 and 2. This is connected with relatively small total amount of LUT units in these projects. As a result we have small amount of LUT units placed along the embedding path (the total size of the digit watermark) and, consequently, too small length of the hash sum field.

Projects 3 and 4 have more amount of LUT units than the ones 1 and 2. However in applying the traditional method the hash sum field size for these projects permits to use only a hash sum obtained with the help of hash function MD5 (the size is 128 bits).

For project 5 the traditional method gives the possibility to use a hash sum obtained with the help of both hash function MD5 (the size is 128 bits) and hash function SHA1 (the size is 160 bits).

The method offered in the given paper provide the effective volume (for the hash sum storage), which equal to the amount of LUT units placed along the embedding path. Thereby the proposed method allows saving a hash sum in the digital watermark for all projects, which participate in the experiment (Table 1). For those projects, for which the traditional method provides the minimum possible size of the hash sum field, the proposed method permits to use a hash sum of the larger size.

V. CONCLUSIONS AND DIRECTIONS OF THE FURTHER RESEARCH

A method of FPGA program code integrity monitoring based on the digital watermark usage is offered in the paper. The method is different from the similar ones existing in the literature with the fact that it does not apply the compression to save and recover the initial information object state at the stage of monitoring. To provide the recovery of initial information object state within the framework of the proposed method the preliminary bringing of information object to a predetermined state, set by steganographic key, is carried out. For performing the integrity monitoring after extracting the digital watermark the repeated bringing of information object to the specified predetermined state is carried out.

The experimental research of the proposed method has shown its efficiency (as compared to the traditional methods) in the point of provision of the effective digital watermark volume sufficient for saving the hash sums obtained with the help of the most widely used hash functions.

We assume that perhaps the usage of the proposed method reduces (as compared to the traditional compression-based methods) the program code information object resistance to stegoanalysis. We assume that perhaps the usage of the proposed method reduces (as compared to the traditional compression-based methods) the program code

information object resistance to steganalysis. Here we mean only the process of detection of the digital watermark presence in FPGA program code. But the question if the information object resistance to steganalysis becomes less (and if it decreases then to what extension) requires extra research. If as a result of this research we come to the conclusions that the resistance really reduces then the technique of the following compromise variant choice is to be created: what is more important – to use a hash sum with more amount of bits (with larger cryptographic secure) or to decrease the probability of detection of the digital watermark in the program code.

Thus the further research of the proposed method is as follows: to study the information object resistance to steganalysis as compared to the traditional methods.

REFERENCES

- [1] I. Habli, R. Hawkins and T. Kelly, "Software safety: relating software assurance and software integrity," *International Journal of Critical Computer-Based Systems*, No 1 (4), pp. 364–383, 2010.
- [2] M. Bishop, *Computer Security*. 2nd edn. USA, Boston: Addison-Wesley 2018.
- [3] W. Stallings, *Cryptography and Network Security: Principles and Practice*. 7th edn. United Kingdom, Harlow: Pearson Education Limited, 2017.
- [4] V. Kharchenko, A. Gorbenko, V. Sklyar and C. Phillips, "Green Computing and Communications in Critical Application Domains: Challenges and Solutions," in 9th International Conference on Digital Technologies (DT2013), pp. 191-197. Zhilina, Slovak Republic, 2013.
- [5] D. Maevsky, A. Bojko, E. Maevskaya, O. Vinakov and L. Shapa, "Internet of things: Hierarchy of smart systems," in 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 2, pp. 821-827, 2017.
- [6] V. Kharchenko, O. Illiashenko, A. Kovalenko, V. Sklyar and A. Boyarchuk, "Security Informed Safety Assessment of NPP I&C Systems: GAP-IMECA Technique," in 22nd International Conference on Nuclear Engineering, pp. 1–9. Prague, Czech Republic, 2014.
- [7] A. Drozd, M. Drozd and V. Antonyuk, "Features of Hidden Fault Detection in Pipeline Components of Safety-Related System," *CEUR Workshop Proceedings*, vol. 1356, pp. 476–485, 2015.
- [8] J. Andina, *FPGAs: Fundamentals, Advanced Features, and Applications in Industrial Electronics*. USA, Boca Raton: CRC Press, 2017.
- [9] W. Vanderbauwhede and K. Benkrid, *High-performance computing using FPGAs*. USA, New-York: Springer, 2016.
- [10] V. Sklyarov, I. Skliarova, A. Barkalov and L. Titarenko, *Synthesis and Optimization of FPGA-Based Systems*. Berlin: Springer, 2014.
- [11] N. Sklavos, R. Chaves, G. Natale, and F. Regazzoni (eds.), *Hardware Security and Trust: Design and Deployment of Integrated Circuits in a Threatened Environment*. Switzerland, Cham: Springer, 2017.
- [12] O. Kehret, A. Walz and A. Sikora, "Integration of Hardware Security Modules into a Deeply Embedded TLS Stack," *International Journal of Computing*, vol. 15, Issue 1, pp. 22-30, 2016.
- [13] J. Vacca, *Computer and information security*. 2nd edn. USA, Waltham: Morgan Kaufmann Publishers, 2013.
- [14] N. Ferguson, B. Schneier and T. Kohno, *Cryptography engineering*. USA, Hoboken: Wiley, 2013.
- [15] J. Katz, *Digital signatures. Advances in Information Security*. USA, New York: Springer, 2010.
- [16] W. Berchtold, M. Schafer and M. Steinebach, "Leakage detection and tracing for databases," in *ACM Information Hiding and Multimedia Security Workshop*, 2013.
- [17] K. Zashcholkin and O. Drozd, "The Detection Method of Probable Areas of Hardware Trojans Location in FPGA-based Components of Safety-Critical Systems," in: IEEE 9th International Conference on Dependable Systems, Services and Technologies DESSERT-2018, pp. 220–225, Kiev, Ukraine, 2018.
- [18] L. Bossuet and L. Torres (eds.), *Foundations of Hardware IP Protection*. USA, New-York: Springer, 2018.
- [19] P. Mishra, S. Bhunia and M. Tehranipoor, *Hardware IP Security and Trust*. USA, New-York: Springer, 2017.
- [20] J. Fridrich, *Steganography in Digital Media*. USA, New York: Cambridge University Press, 2010.
- [21] K. Juneja and S. Bansal, "Frame Selective and Dynamic Pattern Based Model for Effective and Secure Video Watermarking," *International Journal of Computing*, vol. 18, Issue 2, pp. 207-219, 2019.
- [22] M. Arnold, M. Schmucker and S. Wolthusen, *Techniques and Applications of Digital Watermarking and Content Protection*. Boston: Artech House, 2003.
- [23] A. Drozd, M. Drozd and M. Kuznietsov, "Use of Natural LUT Redundancy to Improve Trustworthiness of FPGA Design," *CEUR Workshop Proceedings*, vol. 1614, pp. 322–331, 2016.
- [24] A. Drozd, M. Drozd, O. Martynyuk and M. Kuznietsov, "Improving of a Circuit Checkability and Trustworthiness of Data Processing Results in LUT-based FPGA Components of Safety-Related Systems," *CEUR Workshop Proceedings*, vol. 1844, 654–661, 2017.
- [25] A. Barkalov, L. Titarenko, I. Zeleneva and S. Hrushko, "Implementing on the field programmable gate array of combined finite state machine with counter," in: IEEE 9th International Conference on Dependable Systems, Services and Technologies DESSERT-2018, pp. 247–251. Kiev, Ukraine, 2018.
- [26] K. Zashcholkin and O. Ivanova, "LUT-object integrity monitoring methods based on low impact embedding of digital watermark," in 14th International Conference "Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET-2018)," pp. 519–523. Lviv-Slavske, Ukraine, 2018.
- [27] Ching-Nung Yang, Chia-chen Lin and Chin-chen Chang: *Steganography and Watermarking*. USA New York: Nova Science Publishers, 2013.
- [28] A. Drozd, M. Lobachev, W. Hassonah, "Hardware Check of Arithmetic Devices with Abridged Execution of Operations," in *Proc. European Design and Test Conf, Paris, France*, p. 611, 1996. DOI: 10.1109/EDTC.1996.494375.
- [29] V. Golovko, Y. Savitsky, T. Laopoulos, A. Sachenko, L. Grandinetti, "Technique of learning rate estimation for efficient training of MLP," in *Proc. of the International Joint Conference on Neural Networks, IJCNN-2001, Washington, USA*, pp.323-328, 2001.
- [30] K. Zashcholkin and O. Ivanova, "The control technology of integrity and legitimacy of LUT-oriented information object usage by self-recovering digital watermark," *CEUR Workshop Proceedings*, vol. 1356, pp. 498–506, 2015.
- [31] A. Drozd, S. Antoshchuk, J. Drozd, K. Zashcholkin, M. Drozd, M. Kuznietsov, M. Al-Dhabi and V. Nikul, "Checkable FPGA Design: Energy Consumption, Throughput and Trustworthiness," in: V. Kharchenko, Y. Kondratenko, J. Kacprzyk (eds.) *Green IT Engineering: Social, Business and Industrial Applications, Studies in Systems, Decision and Control*, vol. 171, pp. 73-94. Springer, Heidelberg, 2019.
- [32] A. Sachenko, V. Kochan, V. Turchenko, "Intelligent distributed sensor network," in *Proc. of the IEEE Instrumentation and Measurement Technology Conference IMTC-1998, St. Paul, MN, USA*, pp.60-66, 1998.
- [33] D. Salomon and G. Motta, *Handbook of data compression*, London: Springer, 2010.
- [34] A. Drozd, J. Drozd, S. Antoshchuk, V. Antonyuk, K. Zashcholkin, M. Drozd, O. Titomir, "Green Experiments with FPGA," in V. Kharchenko, Y. Kondratenko, J. Kacprzyk (Eds.), *Green IT Engineering: Components, Networks and Systems Implementation*, vol. 105, Springer, Berlin, 2017, pp. 219-239. DOI: 10.1007/978-3-319-55595-9_11.
- [35] P. Wong, "A public key watermarking for image verification and authentication," in *Proc. IEEE International Conference Image Processing*, pp. 455–459. USA, Chicago, 1998.
- [36] F. Shih, *Digital Watermarking and Steganography: Fundamentals and Techniques*. 2nd edn. USA, Boca Raton: CRC Press, 2017.
- [37] Intel Quartus, <https://www.intel.com/content/www/us/en/software/programmable/quartus-prime/overview.html>, last accessed 2019/07/30.
- [38] Intel Cyclone FPGA series, <https://www.intel.com/content/www/us/en/products/programmable/cyclone-series.html>, last accessed 2019/07/30.
- [39] Intel FPGA Architecture, <https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/wp-01003.pdf>, last accessed 2019/07/30.
- [40] Y. Yang, F. Chen, X. Zhang, J. Yu and P. Zhang, "Research on the Hash Function Structures and its Application," in *International Conference Wireless Personal Communications*, 2016.
- [41] D. Kleidermacher and M. Kleidermacher, *Embedded Systems Security: Practical Methods for Safe and Secure Software and Systems Development*. USA Boston: Newnes, 2012.

Unit Regression Test Selection According To Different Hashing Algorithms

Hakobyan Hovhannes H., Vardumyan Arman V., Kostanyan Harutyun T.
Synopsis Armenia Educational Department
Synopsis Armenia CJSC
Yerevan, Armenia

hhovo@synopsys.com, vardumya@synopsys.com, harutyk@synopsys.com

Abstract – An approach for effective regression test selection is proposed, which minimizes the resource usage and amount of time required for complete testing of new features. Provided are the details of the analysis of hashing algorithms used during implementation in-depth review of the software, together with the results achieved during the testing process.

Keywords – testing, regression test selection, hashing.

I. INTRODUCTION

During the lifetime of every software new revisions and service packs are getting released frequently, among numerous causes the primary ones are bugfixes and enhancements. After any modification, before being delivered to the customer, the software needs to undergo the process known as regression testing, the purpose of which is to ensure that implemented changes did not affect the operability of the final product. This operation can be rightfully considered as one of the vital phases in software development lifecycle. During this step all the issues that were missed before are getting discovered. Unfortunately, most of the time this process is very time consuming and in most cases the deadlines don't give the opportunity to perform the complete validation, thus a dire need for efficient testing mechanism occurs. A number of approaches were elaborated by specialists to achieve the paramount goal of testing time reduction. One of them suggests considering the initial set of test suites, which contains all the low-level test cases that were realized to validate the penultimate version of the software. But instead of going through the whole process of regression testing in order to qualify the new version of the software, all the test cases should be classified to either of 3 main categories. First category includes outdated test cases that are of no use for the updated program, second one contains unnecessary test cases that can be skipped without affecting the quality of the process. The last class specifies the test cases required to be run during the regression testing, as they execute the modified parts of the code. In order to effectively filter out the third category test cases, the proposal is to utilize the principles of Regression Test Selection (RTS) mechanisms. Those are the following:

1. Identification of modified parts of the source code – when a single function inside the program is changed, other modules calling it will be affected as well,
2. Selection of required test cases – this primarily consists of the process of picking the final test cases from the already defined set of test suites.

The scheme of RTS operation is depicted in Fig. 1. This approach of test selection is guaranteed to be secure, as the final set of test cases will be comprised of the initial set of test suites that were already considered as being safe to validate the last but one version of the software. In addition to that the average amount of time spent on regression testing will reduce drastically as total count of final test cases will be lower compared to the initial one.

In addition, unlike other RTS techniques, for instance random selection of certain percent of test cases from the whole set, this approach covers solely the modified portions of the code.

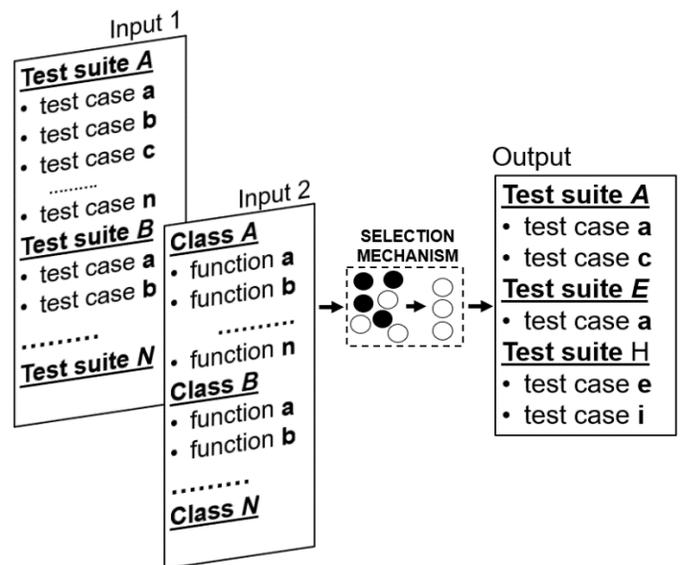


Fig. 1. RTS operation scheme

In order to avoid usage of complicated data structures for storing code in the memory and its further parsing for identification of changes, hashing was used as a part of the proposed mechanism. Due to that the ultimate implementation will not heavily depend on the language being used and will be more versatile and easily adaptable to the requirements.

II. HASHING ALGORITHMS PRINCIPLES

This section illustrates modern hashing methods that were taken into consideration in the scope of the mechanism.

To achieve better performance a number of hashing algorithms were taken into consideration. Those techniques are making use of specific hash functions, the purpose of which is the conversion of a randomly sized input data onto a fixed sized output data. Algorithms analyzed in this paper are MD5 and members of Secure Hashing Algorithms (SHA) family: SHA-1, SHA-2 and SHA-3. There are 4 vital criteria for every hashing algorithm that should be satisfied:

1. It must be capable of calculating mapped values for any sort of data in affordable amount of time,
2. The receiver must not be able to recreate the initial message from its mapped value,
3. Possibility of producing the same hash for differing input data must be minimal, or in ideal not exist at all, this case is known as collision, Table 1 presents the security levels for different algorithms against collisions,
4. Any type of change in the input data, including a tiny one, must cause drastic changes in final hash value, this is known as the avalanche effect.

TABLE I. ALGORITHM TYPES AND SECURITY LEVELS

Algorithm	Security (in bits) against collision
MD5	≤ 18 (collisions detected)
SHA-1	< 63 (collisions detected)
SHA-2	Up to 256
SHA-3	Up to 256

“Message Digest 5” (MD5) generates output hash value equal to 128 bits. Its message processing is conducted on 512 bit blocks and consists of 4 rounds of 16 bit operations. Over time after discovering its numerous vulnerabilities it was deprecated to be used in security systems, but still can be found in applications for data examination against unmeant corruption.

SHA-1 employs the principles introduced in MD5, its output size is equal to 160 bits, and in most applications is represented as a 40-digit hexadecimal number. But unlike MD5

SHA-1 uses 160-bit buffer to process 512 bit blocks and operations during each round are performed on 20 bit blocks. It also was compromised after detection of collisions.

SHA-2 on the contrary to its predecessors, represents a series of hash functions coupled with their digests. Depending on the function being implemented, output values can be 224, 256, 384 or 512 bits long. Each of them differs from the others by the number of shifting operations, values of appended constants and total count of rounds.

SHA-3 is by far the most up-to-date representative of the SHA family. It exploits the principles of sponge construction, where during absorption step, incoming data units are XOR-ed into a subset of the state, which are afterwards rendered as a single block, in follow-up squeezing step, output units are taken from the same subset of the state, interchanged with state transformations.

III. FUNCTION SELECTION PRINCIPLES

This section provides instances of code for better perception of RTS system’s working principles.

Figures presented below represent a tiny part of a program written in Python scripting language. These examples are depicted to explain what types of code modifications can be interpreted as functional changes and for which of them test cases will be chosen to be run during the regression testing.

Sync class contains an initialization method, which creates a window with 2 separate buttons named “Force” and “Manual” which indicate the type of project sync that user wants to perform. Depending on the choice appropriate function calls are made via “clicked.connect” command (Fig. 2). Change made here is the addition of a comment describing the method, such changes are not considered as functional, as they don’t affect the compilation outcome of the code. Thus, test cases related to this function will not be included in final test suite.

```
class Sync(QDialog):
    def __init__(self, parent):
        super(ForceManualSync, self).__init__(parent)

        # Change – Top level layout with buttons
        layout = QGridLayout()

        force_btn = QPushButton("Force")
        layout.addWidget(force_btn, 0, 0)
        force_btn.clicked.connect(partial(self.WhichSyncBtn,
"force"))

        manual_btn = QPushButton("Manual")
        layout.addWidget(manual_btn, 0, 1)
        manual_btn.clicked.connect(partial(self.WhichSyncBtn,
"manual"))

        self.setLayout(layout)
        self.setWindowTitle("Sync Type")
        self.show()
```

Fig. 2. Class definition 1, no functional change

WhichSyncBtn function takes the result submitted by the user and calls functions responsible for either force or manual

syncing (Fig. 3). In case of manual syncing process user is asked to select a specific file containing revision numbers, which will be used as a reference. As no changes were implemented on this method all the test cases related to it will not be classified as retestable.

```
def WhichSyncBtn(self, sync_type):
    self.accept()

    if sync_type == "force":
        self.nd = ForceSyncTrack(self)
    elif sync_type == "manual":
        rev_name = QFileDialog.getOpenFileName(self, "Open file", "~%/s/"
        % getpass.getuser(), "Rev files (*.rev)")

        if len(rev_name) > 0:
            manual_sync_rev = rev_name
            self.nd = ManualSyncTrack(self)

    return
```

Fig. 3. Method definition, no functional change

ManualSyncTrack class opens a separate window containing a progress bar to show the process of syncing by means of percentages (Fig. 4). Modifications made here manipulate the principle of percent calculation and result in depiction of different values compared to the initial version. This is considered as a functional change, as it affects the outcome, consequently test cases applied to this function have to be included in ultimate test suite and run during regression testing.

```
class ManualSyncTrack(QDialog):
    def __init__(self, parent):
        super(ManualSyncTrack, self).__init__(parent)

        layout = QGridLayout()
        self.progress = QProgressBar()
        layout.addWidget(self.progress, 0, 0)

        self.setLayout(layout)
        self.setWindowTitle("Sync Track")
        self.show()

        total_files = int((Popen("cat %s | wc -l" % manual_sync_rev, shell =
        True, stdout = PIPE).communicate())[0])

        background = ManualSyncProcess()
        background.start()
        time.sleep(4)

        current = [0, "syncInfo"]

        while int(current[0]) < total_files:
            current = (Popen("wc -l syncInfo", shell = True, stdout =
            PIPE).communicate())[0].split(" ")
            # percent = (int(current[0]) * 100) / int(total_files)
            percent = (int(current[0]) * 1000) / int(total_files) # Changed
            self.progress.setValue(percent)
            time.sleep(5)
```

Fig. 4. Class definition 2, with functional change

ForceSyncTrack class in a nutshell operates the same way as the ManualSyncTrack, with the distinct difference being the

approach for calculating the value of “total files” variable (Fig. 5). Change made here is the alteration of the variable name – from “current” to “present”. Even though such name shifts don’t affect the functionality of the program, still test cases generated for it should be rerun to assure changes were made in all appropriate places without exception.

```
class ForceSyncTrack(QDialog):
    def __init__(self, parent):
        super(ForceSyncTrack, self).__init__(parent)

        layout = QGridLayout()
        self.progress = QProgressBar()
        layout.addWidget(self.progress, 0, 0)

        self.setLayout(layout)
        self.setWindowTitle("Sync Track")
        self.show()

        total_files = int((Popen("p4 sync -nf %s/... | wc -l" % proj_dir, shell =
        True, stdout = PIPE).communicate())[0])

        background = ForceSyncProcess(proj_dir)
        background.start()
        time.sleep(4)

        # Change – current to present
        present = [0, "syncInfo"]

        while int(present[0]) < total_files:
            present = (Popen("wc -l syncInfo", shell = True, stdout =
            PIPE).communicate())[0].split(" ")
            percent = (int(present[0]) * 100) / int(total_files)
            self.progress.setValue(percent)
            time.sleep(5)
```

Fig. 5. Class definition 3, no functional change

IV. PROPOSED APPROACH

Figure 6 illustrates the structural block scheme of the proposed test selection system. For each of the blocks the required entry information and expected output data are depicted.

In the first phase, as an input argument the source code of the modified software is being used. The purpose is to parse it entirely in order to get the complete hierarchy of classes and children functions or methods used in the program.

The second phase takes 2 input parameters – list of functions retrieved from the previous phase and a hashing algorithm specified by the user or depending on the implementation. The key here is to go over each line of functions’ body codes and deriving a corresponding hash value for it and comparing it with the hash value of the same line of the penultimate version of the code. Some changes though, similar to those discussed in third section, are skipped during the hashing process. The difference in obtained values will mean the presence of distinctions in the code, thus the function was updated and should be passed to the next phase.

The third phase uses the provided list of updated functions together with the set of test suites in order to generate the list of

absolutely essential test cases for regression testing. One major requirement here is that test suites should be provided in the form of a coverage matrix, which is used to map test cases with the functions they were generated for.

In the final phase all the selected test cases are gathered together to form a final test suite to verify the quality of the latest version of the software.

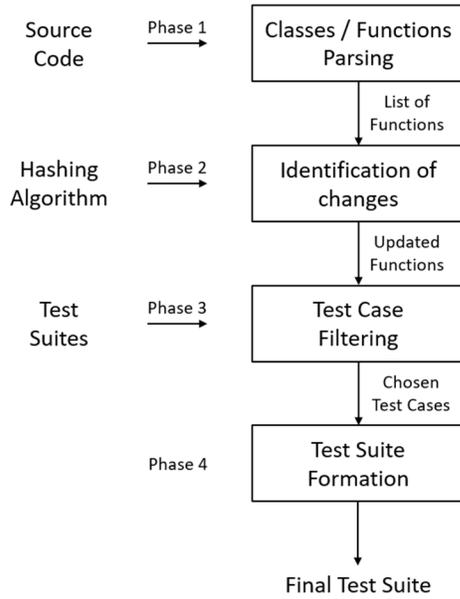


Fig. 6. Block scheme of proposed RTS method

Depending on the syntax specifics of the language for which the source code is being checked, even more occasions of changes can be considered as nonfunctional, thus making the filtering process performed in the 2nd phase more effective. That will result in selection of lesser number of required test cases, and boost the overall performance of the regression testing process.

V. RESULTS

Figure 7 picturizes the average results of the analysis performed on MD5, SHA-1, SHA-2 and SHA-3 hashing algorithms. On the plot X axis corresponds to the number of functions present in the source code, each of them containing on average 1000 lines of code, Y axis corresponds to the overall execution time in seconds. More accurate and detailed results for exact runtimes for a source code with 200 functions can be found in Table 2.

As can be inferred from the data represented on the plot and the table, comparatively fast performing algorithm is the MD5, followed closely by SHA-3. But compared to SHA-3 MD5 is relatively easy implement on different platforms as it's hashing function isn't much complicated. But on the other hand, it's more prone to collisions, even though the appearance relativity of which is very low, especially for the purpose it's being analyzed.

All the tests were conducted on a CPU with Intel i7 7th generation processor with 4 cores, 2.8GHz operating frequency and 16GB RAM. Operating system installed is 64-bits.

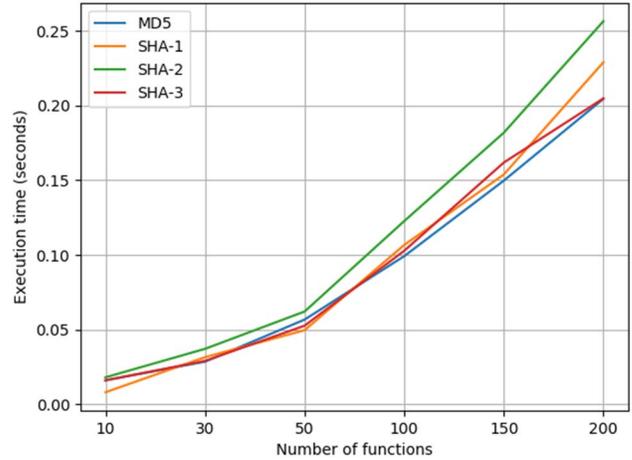


Fig. 7. Relation of execution time to number of functions for algorithms

TABLE II. RUN TIME FOR DIFFERENT ALGORITHMS

Runs \ Alg.	Run 1	Run 2	Run 3	Run 4
MD5	0.300	0.184	0.183	0.215
SHA-1	0.231	0.200	0.184	0.199
SHA-2	0.434	0.247	0.244	0.352
SHA-3	0.251	0.199	0.315	0.316

VI. CONCLUSION

A new test selection mechanism is proposed, which exploits the principles of RTS methods coupled with change identification system and trackability matrixes. This method recommends using hashing instead of other complex parsing techniques. The detailed results of the analysis were presented. Although this approach allows implementation of more sophisticated code preprocessing phases, which can pose as a scope for future research. They can be implemented for effectively filtering out the parts of the code that include changes which are not functional, thus should be emitted from hashing phase.

VII. REFERENCES

- [1] L. Briand, Y. Labiche, and S. He. "Automating regression test selection based on UML designs" Information and Software Technology, Vol 51, No. 1 pp. 16–30, January 2009.
- [2] P. K. Chittimalli. "Regression test selection on system requirements" Conference Paper, January 2008, pp. 87-96.
- [3] Debnath, Santanu, Abir Chattopadhyay, and Subhamoy Dutta. "Brief review on journey of secured hash algorithms." 2017 4th International Conference on Opto-Electronics and Applied Optics (Optronix). IEEE, 2017.

SCOAP-based Directed Random Test Generation for Combinational Circuits

Seyyede Maryam Ghasemy, Maryam Rajabalipanah, Saeideh Sarmadi, Zainalabedin Navabi
School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
{ s_ma_ghasemy, m.rajabalipanah, s.sarmadi, navabi }@ut.ac.ir

Abstract— In this paper, a heuristic method based on a pre-calculated prediction of tests is presented. This method is for combinational test generation and applies to the combinational parts of sequential circuits equally well. The proposed method is Directed Random Test Generation that is based on SCOAP observability and controllability parameters. RTG methods are usually evaluated based on fault coverage that grows exponentially with the number of test vectors. Instead of pure random testing with a massive number of test vectors, fewer test vectors are selected by Directed Random Test Generation, and more faults are detected in the early stages of test generation. As a result, fewer faults remain for the deterministic test stage, which is helpful where online testing is considered. A comparison between random testing and directed random testing is performed on combinational benchmarks, using MATLAB and ModelSim simulation environments. The focus is on the stuck-at fault models in random logic circuits. The simulation results show our method has a sharper increase in fault coverage when compared with pure random test generation.

Keywords—Random Test Generation, Directed Random Test Generation, Fault Coverage, Test Vector, Stuck-at fault, Combinational circuits, SCOAP Parameters.

I. INTRODUCTION

Test vectors are generated for the post-manufacturing test of a digital system. Because of the complexity of digital systems, the size of necessary tests and test quality factors is an important subject. Considering the complexity and expansion of today's circuits, test process which applies all possible input combinations to the faulty circuit model and search for those that produce different output than the good circuit becomes a critical issue.

To simplify the problem discussed above, different test generation techniques have been introduced, among which some deterministic and random methods can optimize the search of test vectors. RTG (Random Test Generation) methods can be classified into three different categories: pure random, constrained random, and directed random. The size and complexity of designs have a high impact on the efficiency of pure random methods. Constrained random test generation method attempts to find the appropriate test vectors that lead to a set of important faults of the circuit under test. On the other hand, a directed random test generation method selects one test to be able to detect a specific fault. Obviously, less effort is needed to reach the same coverage goal using a directed test compared to pure random and constrained-random tests.

Decisions about the number of random tests that can be useful in detecting faults, and expected the number of faults to detect can be made based on how many hard-to-detect faults are in the circuit, and how hard it is to detect them. The usage of SCOAP parameters [1] can be helpful in making some of these decisions.

In today's technology, hardware accelerators have gained significant importance. Most of them include multipliers and adders which are regular logic circuits and some consist of random logic circuits such as encoders, decoders, lookup tables, multiplexers, controllers, etc. Many works have been done on adders and multipliers [2-5]. We propose a directed random test generation technique using heuristics based on SCOAP parameters to limit the search space in random logic circuits. With this heuristics, for the same number of test vectors, fault coverage exponential has a much faster growth than the case of pure random test.

The rest of this paper is organized as follows. The next section discusses related works. Section III presents an overview. A brief introduction to SCOAP parameters given Section IV. Section V describes the proposed method. Experimental results are presented in Section VI and we end the paper with some concluding remarks in Section VII.

II. RELATED WORK

Since our method covers test generation and fault coverage estimation, we review the researches in both categories.

The author in [6], developed a framework for directed test generation, bug localization and bug correction in arithmetic circuits based on the remainder produced by equivalence checking methods. The approach in [7] utilized concolic testing to cover only one single target in RTL models. It used static analysis of distance metric in control flow graphs (CFGs) to guide the search, eventually converging to the target. In [8], the author introduced an automated test generation technique for activating multiple targets, unlike the previous method that focuses on only single faults.

The approach in [9] considers a given test set whose fault coverage is to be calculated. It selects an initial subsequence of test set which is called the training set, then it uses regression analysis to estimate the fault coverage of the test set. This approach accurately estimates the fault coverage by a test set without explicit fault injection.

In [10], a relation between an RT level coverage metric and gate-level fault coverage is represented and it inspired the authors to estimate the fault coverage for the test sets. In [11] and [12], another method using Gate Input Combinations (GIC) is demonstrated. In order to calculate a better fault coverage estimation, the GIC is measured during the logic simulation and partial fault simulation. The authors in [13] use stand-alone fault simulation of modules and some probabilistic analysis to estimate the fault coverage of the entire circuit.

III. OVERVIEW

There are several methods to evaluate the test set generated by random test generation. These evaluations are done based on various parameters of circuits. Similarly, we have developed a heuristic fault coverage estimation method for combinational circuits. In this paper, the single stuck-at fault model has been used. A pre-calculated method is proposed to reach an exponential curve estimating the fault coverage versus the number of test vectors, see Fig. 1. In order to estimate the exponential curve, SCOAP parameters are used. We utilized NetlistGen tool and ModelSim PLI (Programming Language Interface) library [14] to calculate observability and controllability.

Considering SCOAP parameters, faults can be categorized into three groups, consisting of hard to detect faults, easy to detect faults, and some that fall in between these two groups. We assume an exponential growth in the number of faults detected per test vectors. More faults are detected by the early test, that exponentially decays by the later tests. This becomes the estimated direction that we expect from test vectors. The unknowns of the exponential equation are calculated based on hard to detect and easy to detect faults.

IV. SCOAP PARAMETERS

In deterministic or random test generation, controllability and observability measures are used for simplifying the related algorithms. Sandia Controllability/Observability Analysis Program (SCOAP) [1] is a testability measure, the complexity of which grows only linearly with the size of the circuit. SCOAP is based on the topology of the circuit, is a static analysis. It is easy to calculate and provide a good estimate for test generation programs, as well as design for test techniques [15].

SCOAP defines a set of parameters for combinational and sequential controllability and observability measures. The combinational parameters, that are of interest here, have to do with the number of lines that need to be set for controlling and observing a line.

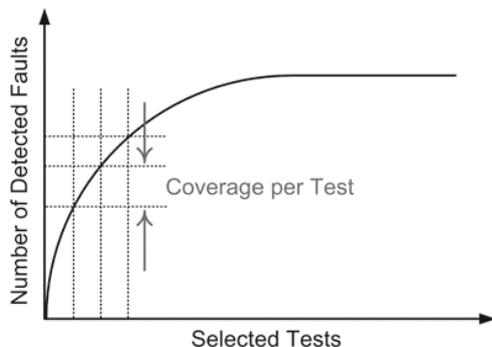


Fig. 1. The exponential growth of detected faults

In SCOAP parameters, lower values mean more controllable and observable, and lines that are more difficult to control and observe have higher SCOAP parameter values.

SCOAP defines a set of parameters for combinational controllability and observability, and another set for sequential. The combinational parameters are $CC0(L)$, $CC1(L)$, and $CB(L)$. $CC0(L)$ and $CC1(L)$ are for Combinational 0-controllability and 1-controllability of line L , respectively. These parameters relates to the number of lines from the primary inputs that have to be traced to put a 0 or a 1 on line L , respectively. The other parameter is $CB(L)$, that defines combinational observability of line L . This parameter relates to the number of lines that have to be traced to observe the value of line L on primary output.

Primary input combinational controllability values are 1 (most controllable) and primary output combinational observability values are 0 (most observable).

Combinational parameters for a circuit line are represented in curly brackets as $\{(CC0,CC1),CB\}$. For calculation of the controllability parameters, initially, $CC0$ and $CC1$ values for lines connected to primary inputs are determined and put in a set of parenthesis in curly brackets. After the determination of controllability values for the primary inputs, SCOAP calculations for the rest of the circuit for logic levels closer to the primary inputs are performed. Therefore, $CC0$ and $CC1$ values for lines leading to the outputs of the circuit are calculated. This procedure continues forward until primary circuit outputs are reached.

Calculation of CB values begins with the primary outputs and moves toward the primary inputs. CB for the primary outputs is 0. CB for lines leading to the inputs are calculated until all lines are examined [15].

As an example, we have used SCOAP parameter calculation to c17 circuit of ISCAS 85, and the results are shown in the diagram of Fig. 2 In the experimental section, several other benchmarks are used for which the SCOAP parameters are calculated in a similar fasion.

V. PROPOSED METHOD

This section describes our method of estimating an exponential for a given circuit. The steps to follow are as shown below.

- Using NetlistGen tool, the gate-level net-list of the circuit is generated.
- We use the PLI function in a Verilog environment for SCOAP calculations. The PLI function takes the netlist as the input and produces SCOAP parameters and circuit levels.
- Equations based on Step 2 outputs are written to estimate the behavior of fault coverage.

In the first step, a number of test vectors are applied to the circuit. We inject a single stuck-at fault in the circuit to make a faulty circuit model, repeatedly. For each fault, a test with the largest amount of covered faults is selected. This provides a good template for estimation of fault coverage.

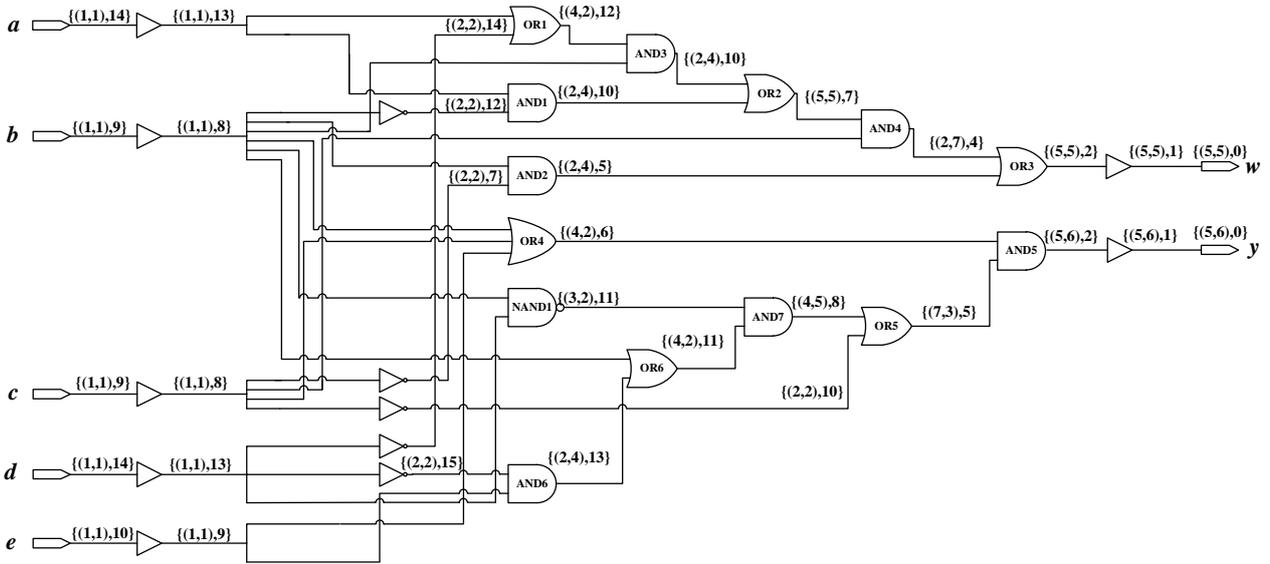


Fig. 2. SCOAP parameters for ISCAS85 c17

According to the behavioral curve of fault coverage with respect to each test vector, an exponential trend of the graph is observed. We use the natural exponential curve to display the mentioned behavior. As the obtained graph cannot be perfectly fitted with a one-phase exponential curve, two-phase exponential fitting is established. This is illustrated in Fig. 3 for c1355.

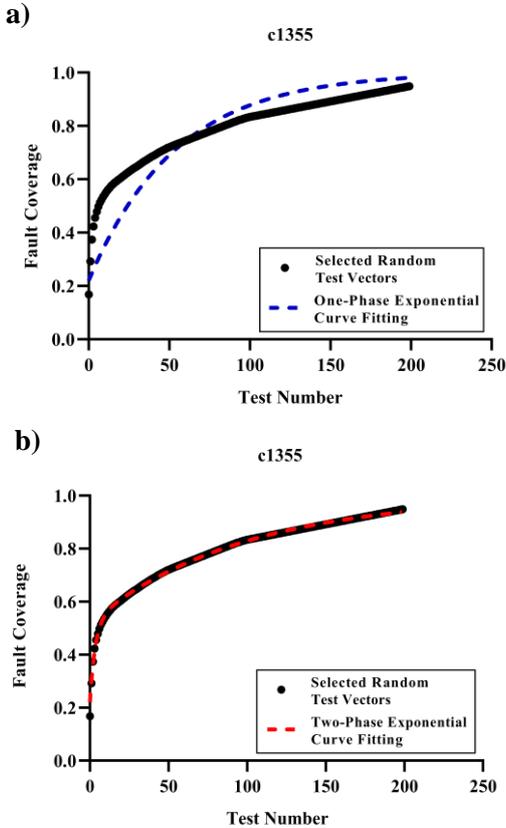


Fig. 3. A comparison between fitting the one and two-phase exponential curves with desired graph

Based on the results of MATLAB and GraphPad Prism, it can be observed that our curve and the two-phase exponential curve match with greater accuracy than the one-phase exponential curve. The estimated curve includes two exponential terms, one of which has a larger growth rate called Fast exponential term, and the other with a lower growth rate called Slow exponential term. We model this two-phase exponential curve with the following equation that is called FC equation.

$$FC = FC_0 + \alpha \times (1 - e^{-K_{Fast} \times t}) + \beta \times (1 - e^{-K_{Slow} \times t}) \quad (1)$$

In this equation, α and β are the coefficients that control the participation rate of faster and slower terms, K_{Fast} and K_{Slow} determine the growth rate of the exponential curve. Moreover, t represents the number of test vectors and FC denotes the Fault Coverage, as the input and output values of (1), respectively.

The summation of CC_0 and CB is used as a measure of stuck-at 0 faults detection capability which we call SUM_0 . In a similar way, CC_1+CB is considered for stuck-at 1 faults as SUM_1 . The standard deviation (δ) for each summation is calculated with respect to the average (μ) for both stuck-at 0 and stuck-at 1 faults. The equation to estimate FC_0 is described in (2).

$$FC_0 = \frac{\text{size}(ETD_0) + \text{size}(ETD_1)}{2 \times \text{number of gates}} \quad (2)$$

In the above equation, ETD_0 and ETD_1 are the set of lines. A line is placed in ETD_0 if its SUM_0 value is less than $\mu_0 - \delta_0$. The same goes for ETD_1 . The numerator of (2) adds the number of faults in ETD_0 and ETD_1 sets.

Equations (3) and (4) describe the sets of ETD_0 and EDT_1 .

$$ETD_0 = \{(SUM_0)_i | (SUM_0)_i < \mu_0 - \delta_0\} \quad (3)$$

$$ETD_1 = \{(SUM_1)_i | (SUM_1)_i < \mu_1 - \delta_1\} \quad (4)$$

We define a criterion to quantitatively gauge the density of each level based on the number of gates. This quantity expresses that those faults placed in the denser parts of a

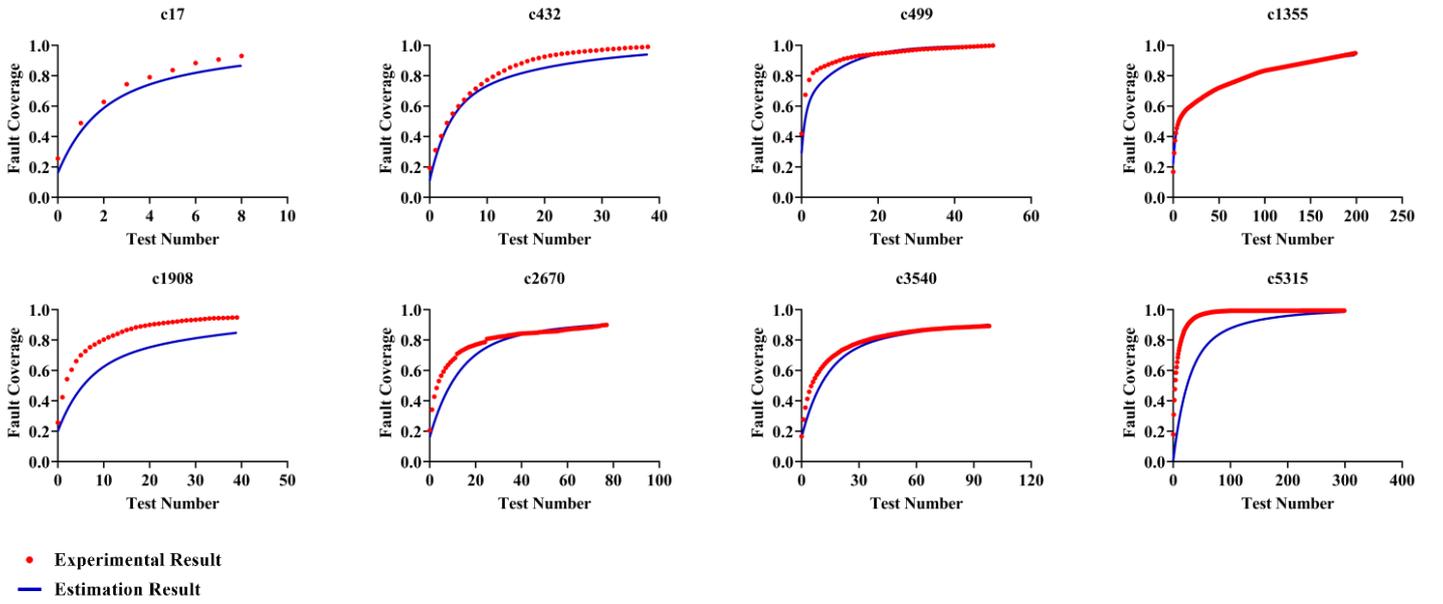


Fig. 4. Comparison between the results of experimental and estimation fault coverage.

circuit are more likely to be hard to detect. Then using (5), the densest level is obtained.

$$\text{The densest level} = \sqrt{\frac{\sum_{i=1}^n l_i^2}{n}} \quad (5)$$

in which, n is the number of gates and l_i refers to the level of i^{th} gate.

In order to indicate how the growth rate of the exponential curve is, we heuristically define the parameter $KSlow$ and $KFast$ as (6) and (8). To do this, the gates are classified into two groups: n_1 and n_2 . The n_1 group covers those gates located between the first and the densest level and n_2 covers the rest.

$$KSlow = \frac{|\sum_{j=1}^{n_1} ((SUM_0)_j + (SUM_1)_j) - \sum_{k=1}^{n_2} ((SUM_0)_k + (SUM_1)_k)|}{2n^2} \quad (6)$$

The parameter $KFast$ can be computed as:

$$KFast = 2.5 \times \frac{\sum_{i=1}^n ((SUM_0)_i + (SUM_1)_i)}{2n^2} \quad (7)$$

The parameters α and β can be calculated as (8) and (9):

$$\alpha = (1 - FC_0) \times (1 - \gamma) \quad (8)$$

$$\beta = (1 - FC_0) \times \gamma \quad (9)$$

where γ

$$\gamma = \frac{1}{\text{number of levels}} \times \text{The densest level} \quad (10)$$

VI. EXPERIMENTAL RESULT

The proposed algorithm has been implemented in Verilog HDL, using ModelSim simulation environment. Experimentation was performed on the ISCAS '85 benchmarks. We develop a desired fault coverage pattern to estimate the proposed FC equation (1). To achieve this goal, we search in a test set for each fault and determine a test vector with the highest fault coverage. It can be observed that the achieved pattern imitates the exponential behavior.

On the other hand, for each benchmark a gate-level netlist is generated in Verilog, using NetlistGen tool. Applying PLI library to the Verilog testbench, the SCOAP parameters, gate circuit levels, and the number of gates are calculated. With the help of GraphPad Prism software, PLI report information and MATLAB, we estimate the FC equation.

Fig. 4 show the accuracy of desired and estimated fault coverage. Table I summarizes the fault coverage improvement by the proposed method in comparison with the pure random test generation.

It can be explicitly realized from Table I that the increase in fault coverage with respect to the number of test vectors is significant. For example, for c499, the number of test vectors produced in the pure random model for fault coverage of 90% is 138, while the number of test vectors in the proposed method with the same fault coverage is 22. As can be seen, the number of test vectors has decreased by a factor of seven.

Another point to be taken from the table is that the maximum of fault coverage by the proposed method cannot be achieved even by generating a large number of test vectors in the pure random method. An example of this fact for c1355, as depicted in Fig. 5.

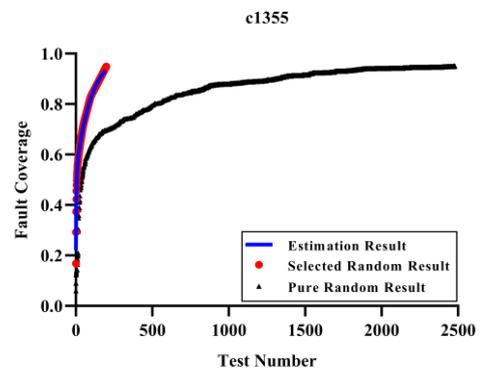


Fig. 5. Comparison between the results of experimental, estimation and pure random fault coverage.

TABLE I. A COMPARISON ON THE NUMBER OF TEST VECTORS BETWEEN PROPOSED DIRECTED METHOD AND PURE RANDOM METHOD FOR DIFFERENT FAULT COVERAGE

FC Benchmarks	Number of Test Vectors													
	50%		60%		70%		80%		85%		90%		95%	
	Proposed	Pure	Proposed	Pure	Proposed	Pure	Proposed	Pure	Proposed	Pure	Proposed	Pure	Proposed	Pure
c17	3	5	4	6	5	7	7	9	9	18	11	34	-	N.O ^a
c432	5	6	7	10	10	17	16	21	21	31	29	65	43	116
c499	2	4	3	5	5	8	9	20	12	26	16	56	22	138
c880	7	7	10	10	16	20	39	48	66	107	106	553	176	6610
c1355	7	41	18	77	47	212	87	516	116	767	157	1312	226	2479
c1908	9	7	12	18	19	69	34	189	49	604	68	1049	104	4847
c2670	10	9	15	18	21	43	33	4515	46	30152	-	N.O	-	N.O
c3540	11	16	16	35	24	100	41	1102	61	3831	98	33448	-	N.O

^a Over 100000 test vectors (Not Observed).

VII. CONCLUSION

By increasing the complexity of the digital circuits, the need for reliable and fast test methods is important. There are several categories of RTG algorithms that are different in mechanisms they use to follow the exponential behavior of fault coverage. According to this, we have proposed a SCOAP-based directed random test generation method. Comparison is done between the result of applying the presented method on ISCAS '85 benchmarks and the result of pure random. Experimental results demonstrate that we gain higher fault coverage than pure random with respect to the number of test vectors.

REFERENCES

- [1] L. H. Goldstein, E. L. Thigpen. "SCOAP: Sandia controllability/observability analysis program." In Papers on Twenty-five years of electronic design automation, pp. 397-403. ACM, 1988.
- [2] A. S. Oyeniran, S. P. Azad, and R. Ubar, "Combined pseudo-exhaustive and deterministic testing of array multipliers," IEEE International Conf. on Automation, Quality and Testing, Robotics, 2018.
- [3] A. S. Oyeniran, S. P. Azad, and R. Ubar, "Parallel pseudo-exhaustive testing of array multipliers with data-controlled segmentation," IEEE International symp on Circuits and Systems (ISCAS), 2018.
- [4] B. Y. Ye, P. Y. Yeh, S. Y. Kuo, and I. Y. Chen, "Scalable and bijective cells for C-testable iterative logic array architectures," IET Circuits, Devices & Systems, vol. 3(4), pp. 172-181, August 2009.
- [5] L. Sekanina, "Design and analysis of a new self-testing adder which utilizes polymorphic gates." In 2007 IEEE Design and Diagnostics of Electronic Circuits and Systems, pp. 1-4. IEEE, 2007.
- [6] F. Farahmandi, P. Mishra, "Automated test generation for debugging multiple bugs in arithmetic circuits." IEEE Transactions on Computers 68, no. 2 (2018): 182-197.
- [7] A. Alif, F. Farahmandi, and P. Mishra, "Directed test generation using concolic testing on RTL models." In 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 1538-1543. IEEE, 2018.
- [8] Y. Lyu, A. Ahmed, and P. Mishra, "Automated activation of multiple targets in RTL models using concolic testing." In 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 354-359. IEEE, 2019.
- [9] P. K. Javvaji, S. Tragoudas, and G. Kondapuram, "Scalable Fault Coverage Estimation of Sequential Circuits without Fault Injection." In 2018 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5. IEEE, 2018.
- [10] S. Park, L. Chen, P. K. Parvathala, S. Patil, and I. Pomeranz, "A functional coverage metric for estimating the gate-level fault coverage of functional tests." In 2006 IEEE International Test Conference, pp. 1-10. IEEE, 2006.
- [11] S. Mirkhani, and J. A. Abraham, "Eagle: A regression model for fault coverage estimation using a simulation based metric." In 2014 International Test Conference, pp. 1-10. IEEE, 2014.
- [12] S. Mirkhani, and J. A. Abraham, "Fast evaluation of test vector sets using a simulation-based statistical metric." In 2014 IEEE 32nd VLSI Test Symposium (VTS), pp. 1-6. IEEE, 2014.
- [13] S. Mirkhani, J. A. Abraham, T. Vo, H. Jun, and B. Eklow, "FALCON: Rapid statistical fault coverage estimation for complex designs." In 2012 IEEE International Test Conference, pp. 1-10. IEEE, 2012.
- [14] A. Kamran, N. Nemati, S. S. Kohan, and Z. Navabi, "Virtual tester development using HDL/PLI." In 2010 East-West Design & Test Symposium (EWDTS), pp. 412-415. IEEE, 2010.
- [15] Z. Navabi, "Digital system test and testable design." E-ISBN (2011): 97814419-97875485.

OR2-NoC: OFFLINE ROBUST ROUTING ALGORITHM FOR 2-D MESH NoCs ARCHITECTURES

Arezoobeheshti soofian
Department of Electrical and Computer
Engineering
University of Tabriz
Tabriz, Iran
arezoobeheshti93@ms.tabrizu.ac.ir

Mina Zolfy lighvan
Department of Electrical and Computer
Engineering
University of Tabriz
Tabriz, Iran
mzolfy@tabrizu.ac.ir

Zahra Eghbali
Department of Electrical and Computer
Engineering
University of Tabriz
Tabriz, Iran
zahra.eghbali14@yahoo.com

Abstract— This paper presents the routing algorithm that tolerate multiple faulty routers and multiple faulty channels for 2-D mesh Network-on-Chips (NoC) and it has a high fault tolerance. The proposed approach always finds a path between two healthy nodes. This algorithm is implemented on the Distributed Scalable Predictable Interconnect Network (DSPIN) architecture. The proposed method is based on the routing table which filled at the beginning of the circuit operation and will be updated once a few times. Experiment result show that the proposed method has high reliability and less packet latency.

Keywords— NoC, Routing, reliability, packet latency, fault tolerance

I. INTRODUCTION (HEADING 1)

The increasing the numbers in of processor in current system-on-chip (SoC), requires more complex communication fabric for providing the on chip connection requirements. In the early SoC designs point-to-point communication structure was common place. As the complexity of SoC, increases overcoming the constraint such as reusability, bandwidth limitation, area overhead, high design time, power consumption and efficiency becomes more challengeable in traditional communication architecture. The challenges leads SoC designers and developers to another modular on chip communication architecture name Network-on-Chip (NoC) like any other digital design, test process is required to detect any probable detect in NoCs[1, 2]. NoCs require some algorithms for performing the on chip communications where routing algorithm is one of them. Routing algorithm determines the packet route between the source and destination and with the purpose select the shortest path while prevent the deadlock, increase fault tolerance, avoidency congestion.

In recent years, different fault-tolerant routing algorithms have been proposed to tolerate faulty components in 2D mesh NoCs. These algorithms use either VCs or turn model to avoid deadlock conditions. In addition, there are some proposed methods that use memory or table-based routing [3]. The many routing algorithms have been proposed for parallel computers that use three or four virtual channels [4-10].

The OR2-NoC algorithm proposed in this paper avoids live lock while keeping a path between any source and destination, without use any virtual channel. This method routing in a NoC with supports multiple faulty communication channels and multiple faulty routers.

The organization of this paper is as follow after this introduction, Section 2 presents the related work then section 3 describes the proposed routing algorithm after that Section 4 presents the experiment results and section 5 concludes this paper.

II. RELATED WORK

There are many routing algorithms with fault tolerance capability for 2-D NoCs. Some algorithms use turn models, such as West-First (WF), North-last (NL), Negative-First (NF).

In [11], spare wires are used to pass through the faulty links, with a high hardware cost. In [3] a hierarchical routing algorithm is proposed which tolerates multiple faulty links and routers. This method is presented in three levels and tolerates one to several faulty channels and routers. The first level proposed for tolerate single faulty channel, but other levels are multi-fault tolerant. This algorithm is designed to tolerate faulty channels based on the XY routing algorithm in which a packet is first routed in the X dimension and then in the Y dimension to achieve the destination. For fault-tolerance, the reconfiguration has been used and a new deterministic routing algorithm is used for all routers on a cycle-free contour around a faulty channel to use new unique paths instead of the broken paths. It is more useful for NoC applications that require somehow a degree of reliability but not much cost [3]. The authors of [12] construct their method based on faulty link tolerance and then extend it in order to tolerate faulty routers and regions by modeling each faulty router by its four surrounding links which are assumed faulty [12].

In routing based on OR2-NoC algorithm, path between two nodes are determined prior to the chip start, thus preventing from

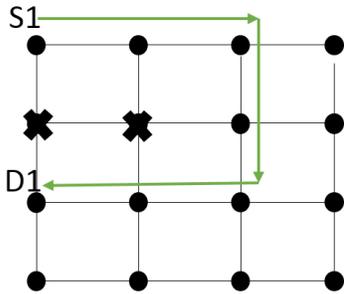


Fig. 1. Routing by Offline algorithm

livelock. This method tolerate multi faulty channels and multi faulty cores and it will always find a path between two healthy nodes and the chosen route between two nodes is the shortest path.

III. PROPOSED METHOD

Using OR2-NoC routing algorithm, each node has its own routing table and use it as lookup table for packet routing. Although in functional mode, these table are distributed our the chip and each table is located on corresponding nodes, but their generation is centralized. They are all generated at the first step of offline routing in the chip root and after that are sent to the nodes. Chip root is one of the nodes on the NoC that is selected in test mode based on some constraints.

In the of this article the process of completing routing table in pointed as root routing. The root routing is done through one of the three following stages. In first stage XY based route in examined in the case of XY route existence, routing table is set. Otherwise go into second stage. In second stage YX route existence in check. Again in the case of successes corresponding row of routing table is set. Otherwise stage three is required. In stage three, routing table of the source neighbors are used to complete the routing table. This trend continues until all undefective. Nodes find a path to each other. And the routing table will complete all the safe nodes.

As the base fault detection method in this paper the mechanism of [13] is used. In this approach, the faulty component are deactivated by means of a BIST mechanism. The presented embedded distributed collaborative configuration infrastructure (DCCI) firmware is executed to probe and explore the NoC, and finally build a temporary communication tree connecting all the operational components [13]. Using this tree's data, we identify the usable nodes and the safe nodes communicate their channels via this tree, we do the routing.

A. Offline Routing Algorithm

In the proposed method, the diagnosis of the defective components is carried out by the solution presented in [13]. The architecture used in implementing approach method is DSPIN (Distributed Scalable Predictable Interconnect Network) [14]. In the proposed method in [13], defective components are detected and deactivated by the BIST mechanism. The presented embedded distributed collaborative configuration infrastructure (DCCI) firmware is executed to probe and

explore the NoC, and finally build a temporary communication tree connecting all the operational components [13]. After detecting the defective parts, the active nodes sent the status of their channels to the root of the tree (root of tree specified in the chip testing phase). Routing is done offline at the root of the tree, since the chip operation phase has not yet begun. In this algorithm, routing table is considered for each node in which the number rows equal to the number of nodes on the chip at the number columns is four.

In proposed offline routing algorithm, first the route between the source and destination is analyzed using the XY algorithm, if the path is free of any faulty component, the routing table is filled in the direction given by the XY algorithm. If not, the path between the source and the destination is checked by the YX algorithm, and if the path not include any defective component, the routing table will be filled in the proposed direction of the YX algorithm. If the path still missing, the algorithm enters the stage where there are some safe nodes that have not found a path each other.

In this stage each other node who failed to find a route to a certain destination looks at the routing table of its neighbors, if any neighbor has a route to this destination, the algorithm fills the table with the table of neighbor node .If multiple neighbors includes a route to the destination at the same time, the shortest path will be selected. "Fig. 1" shows an example of the routing done by using this algorithm and "Fig. 2" shows the algorithm pseudo code. In this code, the XY and YX algorithm returns a value equal to one, when a path exists without a defective component. If $Y_d < Y_s$, YX algorithm returns 4. Otherwise If $Y_s < Y_d$ YX algorithm returns 0. "TABLE I" shows the variable definition of the a pseudo-code presented in "Fig. 2".

B. Reliability Analysis

In this section, the reliability of a NoC is computed using the formulas presented in [3]. The reliability of achieved by multiplying the reliability of all impossible paths between them. This formula is shown in state (1). It should be noted that if any path reliability in the result rises to a power greater than one, this power should be replaced by one. In (1), m is source and n is destination.

$$R_{NoC} = \prod PR_{m \rightarrow n} \quad (1)$$

For computing the $PR_{m \rightarrow n}$ of formula (1), formula (2) is used. In this formula, PL , R_{L_i} and R_{S_j} are respectively path Length (the number of links on a path), link and router reliability.

TABLE I. PSEUDO-CODE VARIABLE DESIGN

RT	Routing Table
C	Current node
N	Neighbor node
n	North
s	South
D	Destination
L	Link
d	Direction

Algorithm Off Line

Begin

1. From all safe nodes to all safe nodes

Call XY

If XY return 1

Fill the Routing Table

else

Call YX

If YX return 1

Fill the Routing Table

2. **for** the nodes that are not filled out from the routing table

While as long as all the safe lines of the routing table have not been filled

If YX returns 0

If $(\exists N, RT_N(D, :) == 1) \&\& (L(C, N) == 1) \&\& (RT_N(D, d_c) != 1)$

If $RT_{N_n}(D, :) == 1 \&\& L(C, N_n) == 1$

$RT_C(D, n) = 1$

Else

$RT_C(D, d_N) = 1$

Else

Continue

elseif YX returns 4

If $(\exists N, RT_N(D, :) == 1) \&\& (L(C, N) == 1) \&\& (RT_N(D, d_c) != 1)$

If $RT_{N_s}(D, :) == 1 \&\& L(C, N_s) == 1$

$RT_C(D, s) = 1$

Else

$RT_C(D, d_N) = 1$

Else

Continue

else

If $(\exists N, RT_N(D, :) == 1) \&\& (L(C, N) == 1) \&\& (RT_N(D, d_c) != 1)$

$RT_C(D, d_N) = 1$

Else

Continue

End

Fig. 2. Pseudo-code offline routing algorithm.

$$PR_{m \rightarrow n} = \prod_{i=1}^{PL} R_{L_i} \times \prod_{j=1}^{PL+1} R_{S_j} \quad (2)$$

The computation performed to calculate the reliability of source – destination connection is shown “Fig. 3”. The path reliability for each source-destination pair will have one term according to (2) if the routing algorithm is XY. But, for a fault-tolerant method the path reliability should have more terms. We assume that router reliabilities equal one. The formulas (3) shows the reliability of the XY algorithm and (4) shows the reliability of the OR2 algorithm.

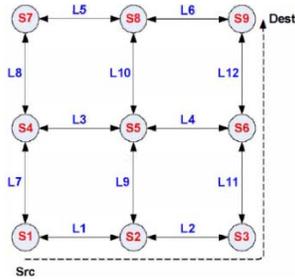


Fig. 3. A 3x3 network with named components

$$PR_{src \rightarrow dest, XY} = R_{L_1} \times R_{L_2} \times R_{L_{11}} \times R_{L_{12}} \quad (3)$$

$$PR_{src \rightarrow dest, Off Line} = PR_{src \rightarrow dest, XY} + 4(1 - R_L)R_L^4 +$$

$$(1 - R_{L_1}) \cdot (1 - R_{L_8}) \cdot R_{L_7} \cdot R_{L_3} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_1}) \cdot (1 - R_{L_5}) \cdot R_{L_7} \cdot R_{L_3} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_1}) \cdot (1 - R_{L_6}) \cdot R_{L_7} \cdot R_{L_3} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_2}) \cdot (1 - R_{L_7}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_2}) \cdot (1 - R_{L_8}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_2}) \cdot (1 - R_{L_5}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_2}) \cdot (1 - R_{L_6}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_{11}}) \cdot (1 - R_{L_7}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_{11}}) \cdot (1 - R_{L_8}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_{11}}) \cdot (1 - R_{L_5}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_{11}}) \cdot (1 - R_{L_6}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_4} \cdot R_{L_{12}} +$$

$$(1 - R_{L_{12}}) \cdot (1 - R_{L_7}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_{10}} \cdot R_{L_6} +$$

$$(1 - R_{L_{12}}) \cdot (1 - R_{L_8}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_{10}} \cdot R_{L_6} +$$

$$(1 - R_{L_{12}}) \cdot (1 - R_{L_6}) \cdot R_{L_1} \cdot R_{L_9} \cdot R_{L_{10}} \cdot R_{L_6} =$$

$$19R_L^4 - 32R_L^5 + 14R_L^6 \quad (4)$$

“TABLE II” shows path reliability of four routing algorithms assuming three different reliability value (RL) for the between neighbor nodes [3].

IV. EXPERIMENT RESULT

All of the implementations and simulation have been done MATLAB environment.

A. Routing Phase

By implementing functions in MATLAB programming environment, the execution time of offline routing is calculated for mesh 4×4, 8×8 and 16×16 with one to three broken routers and one to three broken communication channels, which are randomly injected into the network and then calculated the execution time of this stage is calculated.

“TABLE III” shows the mean of the routing phase with the presence of the defective components.

TABLE II. DIFFERENT RELIABILITIES OBTAINED BY DIFFERENT METHOD FOR PATH SHOWN IN FIG.3

Routing Method	Path Reliability		
	$R_1 = 0.9$	$R_1 = 0.95$	$R_1 = 0.99$
XY	0.66	0.814	0.961
FT-XY	0.91	0.973	0.9988
RDR1	0.95	0.986	0.9994
OR2	1	1	1

TABLE III. MEAN OF ROUTING PHASE WITH THE PRESENCE OF THE DEFECTIVE COMPONENTS (FR: FAULTY ROUTER, FC: FAULTY CHANNEL)

Mesh Size and Faults	Average Duration in Seconds
4×4, 1 FR	0.0143 s
4×4, 2 FR	0.0211 s
4×4, 3 FR	0.0464 s
4×4, 1 FC	0.0146 s
4×4, 2 FC	0.0149 s
4×4, 3 FC	0.0167 s
8×8, 1 FR	0.1048 s
8×8, 2 FR	0.111 s
8×8, 3 FR	0.1918 s
8×8, 1 FC	0.0874 s
8×8, 2 FC	0.0997 s
8×8, 3 FC	0.1123 s
16×16, 1 FR	2.0548 s
16×16, 2 FR	2.0963 s
16×16, 3 FR	2.129 s
16×16, 1 FC	2.1387 s
16×16, 2 FC	2.17 s
16×16, 3 FC	2.1741 s

TABLE IV. MEMORY RATE INVOLVED IN EACH NODE

Mesh size	Memory rate involved in bit
4×4	16×4 b
8×8	64×4 b
16×16	256×4 b

TABLE V. AVERAGE ROUTING TIME IN ROOT WITH BY THE NUMBER OF FAILURES 6 % OF ALL ROUTERS AND 12.5 % OF ALL ROUTERS

NoC Architecture	4×4	8×8	16×16
No. of faulty router (6 % of all routers)	1	4	15
Average routing time in root (second)	0.0142 s	0.3681 s	2.2744 s
No. of faulty router (12.5 % of all routers)	2	8	32
Average routing time in root (second)	0.0231 s	0.6134 s	3.0249 s

B. Storage requirement

In each node, the routing table involves the amount of memory that depends on the size of mesh. As the network size is larger, the amount of memory involved by the routing table will also be greater, because it is checked from each node to all nodes in the path. “TABLE IV” shows memory rate involved in each node.

C. Packet Latency

By implementing MAIN_RAFT algorithm and Offline algorithm in MATLAB programming environment, we investigated the average packet latency when there is no congestion in network. We have calculated the mean of packet latency in the last 50 times, for network 4×4, 8×8 and 16×16 with the existence of four faulty communication channels that were randomly injected into the network. “Fig. 4” shows this calculations.

As the “Fig. 4” shows, packet latency for Offline Robust routing algorithm is less than packet latency for the MAIN_RAFT algorithm.

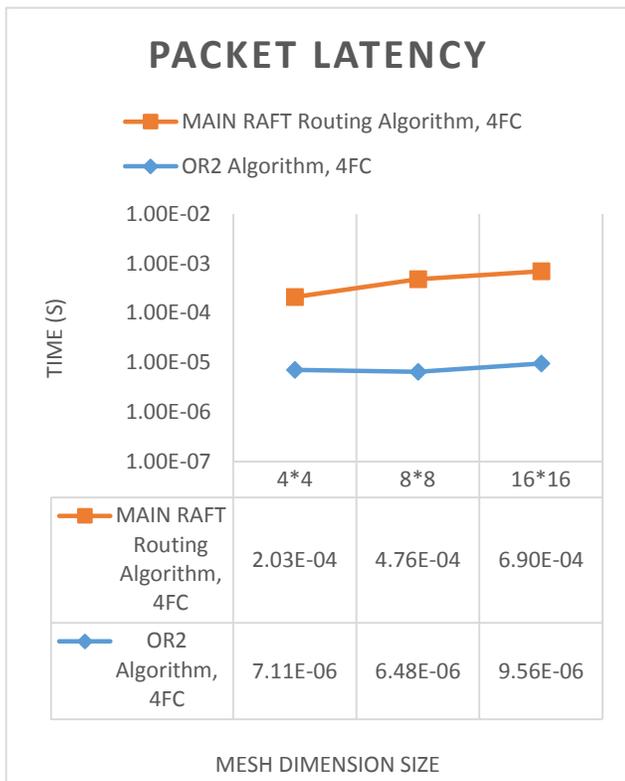


Fig. 4. packet latency

D. Fault Tolerance

The proposed algorithm was tested on 2-D mesh NoC with sizes 4×4, 8×8, 16×16 despite the failure of 6 % of the cores and despite the failure of 12.5 % of the cores and the routing time was measured at the root. “TABLE V” shows the results of these experiments.

V. CONCLUSION

In this paper, a routing algorithm for 2-D mesh NoCs is proposed which is capable of routing, with different types of faulty components including a black holes, multiple faulty link and multiple faulty router. This method uses the routing table, which are filled in the routing phase. Experiment results indicate that this algorithm has high reliability compared to other stated algorithms in this paper and is less packet latency

and always finds a path between the source and destination. The comparison of the algorithms in this article shows that this algorithm reduce the packet latency compared to the RAFT algorithms, and is more reliable than the RDR algorithm.

REFERENCES

- [1] Bjerregaard, T. and S. Mahadevan, *A survey of research and practices of network-on-chip*. ACM Computing Surveys (CSUR), 2006. **38**(1): p. 1.
- [2] Pasricha, S. and N. Dutt, *On-chip communication architectures: system on chip interconnect*. 2010: Morgan Kaufmann.
- [3] Valinataj, M., P. Liljeberg, and J. Plosila. *A fault-tolerant and hierarchical routing algorithm for NoC architectures*. in *NORCHIP, 2011*. 2011. IEEE.
- [4] Boppana, R.V. and S. Chalasani, *Fault-tolerant wormhole routing algorithms for mesh networks*. IEEE Transactions on Computers, 1995. **44**(7): p. 848-864.
- [5] Sui, P.-H. and S.-D. Wang, *An improved algorithm for fault-tolerant wormhole routing in meshes*. IEEE Transactions on Computers, 1997(9): p. 1040-1042.
- [6] Park, S., J.-H. Youn, and B. Bose. *Fault-tolerant wormhole routing algorithms in meshes in the presence of concave faults*. in *Parallel and Distributed Processing Symposium, 2000. IPDPS 2000. Proceedings. 14th International*. 2000. IEEE.
- [7] Xu, Y., J. Zhou, and S. Liu. *Research and analysis of routing algorithms for NoC*. in *Computer Research and Development (ICCRD), 2011 3rd International Conference on*. 2011. IEEE.
- [8] Albughdar, M. and A. Mahmood. *Maximally Adaptive, Deadlock-Free Routing in Spidergon-Domut Network for Large Multicore NOCs*. in *2015 14th International Symposium on Parallel and Distributed Computing*. 2015. IEEE.
- [9] Nayak, C.K., S. Das, and H.S. Behera, *Hierarchical Agents Based Fault-Tolerant and Congestion-Aware Routing for NoC*, in *Computational Intelligence in Data Mining-Volume 3*. 2015, Springer. p. 705-714.
- [10] Khan, G.N. and S. Chui. *Congestion aware routing for on-chip communication in noc systems*. in *Conference on Complex, Intelligent, and Software Intensive Systems*. 2017. Springer.
- [11] Lehtonen, T., P. Liljeberg, and J. Plosila, *Online reconfigurable self-timed links for fault tolerant NoC*. VLSI design, 2007. **2007**.
- [12] Valinataj, M., et al., *A reconfigurable and adaptive routing method for fault-tolerant mesh-based networks-on-chip*. AEU-International Journal of Electronics and Communications, 2011. **65**(7): p. 630-640.
- [13] Zhang, Z., et al., *On-the-field test and configuration infrastructure for 2-D-mesh NoCs in shared-memory many-core architectures*. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2014. **22**(6): p. 1364-1376.
- [14] Miro-Panades, I., et al. *Physical Implementation of the DSPIN Network-on-Chip in the FAUST Architecture*. in *Proceedings of the Second ACM/IEEE International Symposium on Networks-on-Chip*. 2008. IEEE Computer Society.

Simulation of Nodes and Blocks of Matching Processor of the Parallel Dataflow Computing System "Buran"

Nikolay Levchenko
*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
nick@ippm.ru

Anatoly Okunev
*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
oku@ippm.ru

Dmitry Zmejjev
*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
zmejjev@ippm.ru

Abstract — For the parallel dataflow computing system (PDCS) "Buran", depending on the specific application, it is possible to customize the operation of the computational core and the system as a whole to efficiently perform tasks. The article describes the simulation of implementation variants of the matching processor and its individual nodes and blocks. The studies were conducted on the behavioral cycle-accurate simulator and the emulator of the PDCS. These program models tested the correct functioning of the developed sets of nodes and blocks, tested new types of tokens and evaluated the effectiveness of the application of new architectural solutions. The experiments performed demonstrate the speedup of task execution in comparison with the basic variant of the matching processor.

Keywords — set of nodes and blocks, matching processor, behavioral cycle-accurate simulator, computational core

I. INTRODUCTION

The computing model, the application of which may be relevant in the development of supercomputers and other computing systems, is the original dataflow computing model with a dynamically formed context [1]. IPPM RAS is working on the development of this computing model and the creation of its implementation - the architecture of the parallel dataflow computing system (PDCS) "Buran".

This computing model differs significantly from the classical dataflow computing model, which was originally proposed in the 1980s [2–5]. At present, this computing model is experiencing its renaissance [6–9].

The PDCS "Buran" is a multi-core scalable computing system. The standard computational core of this system includes an execution unit, a matching processor, a token commutator, and a hash block. Between the cores in the system, information units are transmitted, representing messages (in the form of tokens), containing an operand, a set of service fields, and a key.

The matching processor, which is part of the PDCS computation core, may consist of nodes and blocks of different functionality. Depending on the range of tasks to be solved, it is possible to customize the computing system (using combinations of different nodes and blocks), so that the execution of a task will be most effective.

The article [10] described approaches to the design of sets of nodes and blocks of the matching processor of the computational core. To implement these approaches, it is

necessary to analyze the algorithms of the tasks, explore various options for implementing sets of nodes and blocks of the matching processor, develop criteria for evaluating these sets, and create programs for testing and simulating different sets of nodes and blocks of the matching processor. The article [10] also described the requirements and specifics of the work of some sets of nodes and blocks. Changing the architecture of the matching processor, optimizing the composition of its nodes and blocks, it is possible to achieve much more efficient execution of tasks of various classes, regulate energy consumption and reduce the equipment footprint.

The goal of this work is to describe the process of simulating various versions of a set of nodes and blocks included in the matching processor (MP) and the results of their simulation. The studies were conducted on the behavioral cycle-accurate simulator and on the PDCS emulator for cluster supercomputers. The scientific novelty of the article is in the analysis of the work functionality of the sets of MP nodes and blocks, in the test of new token types, and in the evaluation of the effectiveness of the application of new architectural solutions for specific tasks.

The relevance of the work is in the fact that this work was carried out as part of a project to create a supercomputer with non-traditional architecture.

II. IMPLEMENTATION OF THE MATCHING PROCESSOR ON THE BEHAVIORAL CYCLE-ACCURATE SIMULATOR AND ON THE EMULATOR

The behavioral cycle-accurate simulator (BCAS) is a software complex developed as part of the PDCS project. The BCAS includes tools that allow the design of the PDCS architecture from a developed set of nodes and blocks, and to simulate the execution of a dataflow program on the resulting architecture. The BCAS uses discrete-event simulation - the processing of the chronological sequence of events generated by objects. All these objects are interconnected through the mechanism of input-output ports, through which data is transferred from one object to another. The sequence of events processing is determined by the time of initialization of data transfer from one object to another. Time is calculated in conditional cycles. Each object in the course of its work changes the internal time parameter and attaches its value to each data transfer sent to another object or group of objects that are connected to one of the output ports of this object. The object receiving this data changes the event's

time in accordance with the value of its internal time parameter and returns to the sender a new value, thereby ensuring the processing of delays in the transmission and reception of data between objects. Thus, the whole process of simulation is based on changing the time parameters of interconnected objects and synchronizing these times through the global time pool, in which the next object for processing is selected.

The selected simulation mechanism allows determining the detalization degree of nodes and blocks of the PDCS matching processor. The operation of the computational module or the core, and, if necessary, the operation of all its components (from complex nodes to individual registers).

The basic set of nodes and blocks of the MP (Fig. 1) is implemented within the framework of a complex computational module as separate functions (for each block and node), the execution or nonexecution of which is regulated by changing the object's parameters. In the BCAS, the researcher, through the parameters of the objects, sets the algorithm for the object operation, as well as the value by which the time of the object increases when each individual node and block is executed.

The MP is a computational process control device in the PDCS that supports matching of the task token keys, provides all the principles of the dataflow computing model, and prevents problems of overflow or underload of hardware resources in the computation process.

The basic set (Fig. 1) of the matching processor includes: content addressable memory of keys (CAMK), response processing block, tokens and packets generation blocks, node for recording and deleting tokens, a node for generating free addresses, and a token memory.

The MP stores the keys with masks, fixes the recorded keys in the vacant register, generates the addresses of free cells to write the keys, and also finds matches with the keys stored in it.

Using MP allows extracting from the task all the parallelism that exists in the algorithm of its solution. That allows efficiently loading the available hardware resources of the computing system and achieving a high degree of actual tasks scaling on systems consisting of hundreds of thousands of computational cores.

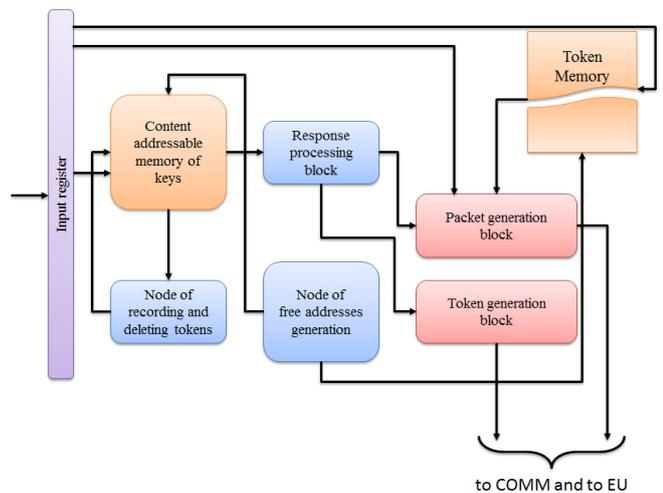


Fig. 1. Basic set of nodes and blocks of the matching processor (COMM – commutator; EU – execution unit)

When simulating the MP, the following sequence of actions (functions) is performed, each of which relates to its separate block:

- The function of receiving the token and writing it to the input register (the token that enters the MP input). The operation of the “input register” is simulated using the “Token reception time” parameter.
- The function of matching input token keys with keys of tokens stored in memory. This function simulates the operation of the block “Content addressable memory of keys”. In the process of work (function execution), the key of the “input” token is compared with each key recorded in the CAMK. As a result of the comparison, a matching list is compiled. The functioning of the block depends on the parameter “Matching time”, the value of which is added when processing each incoming token.
- The function of processing matchings. The work of the "Response processing" block is simulated. The function is executed for each item in the matching list. Codes of operations, multiplicities and other parameters of tokens are analyzed, on the basis of which requests for the execution of subsequent functions are formed. Also during the analysis, the token multiplicity changes — the token parameter that determines the number of matchings to be processed. The operation of this block is determined by the value of the parameter “Matchings processing time”, which is added during the processing of each new element of the matching list.
- The function of recording and deleting tokens. The simulated node is triggered (the function is called) when it is required to record the input token or to delete the token already stored in the CAMK. Record and delete time is determined by the “Record and delete tokens time” parameter. This parameter changes the operation time of the “Free addresses generation” node and the “Recording and deleting tokens” node. The first node is designed to generate a free address in the CAMK and the token memory. The second is for recording a token by the generated address or removing it from the CAMK and the token memory.
- The function of packet generation. The operation of the “Packet generation” block is simulated. The block is activated at the request “generate a packet” and its operation is determined by the “Packet generation time” parameter, the value of which is added each time the function is called.
- The function of the token generation. The operation of the "Token generation" block is simulated. The block is activated at the request “generate a token” and its operation is determined by the “Token generation time” parameter, the value of which is added each time the function is called.

In addition to the represented parameters that determine the operation time of each of the MP blocks, there are parameters that determine the amount of the CAMK and the token memory. Since BCAS is used for study the dataflow systems at the register level and the composition of the token fields and the dimension of these fields can vary, the

measurement unit of the "CAMK size" parameter is a token, and of the "Token memory size" parameter is a 64-bit word. Determining the amount of the CAMK by specifying the maximum number of tokens that can be simultaneously stored in it, simplified the analysis of simulation statistics and assessing the efficiency of passing tasks. Using 64-bit words to determine the maximum amount of the token memory made it possible to simulate the fragmentation of the token memory when standard and multi-input tokens are used in tasks simultaneously. This allows evaluating the effectiveness of the use of multi-input tokens and their impact on the ratio of the amount of the CAMK and the token memory in solving specific tasks.

A similar approach (dividing the work of the PDCS computational core into individual functions responsible for the operation of specific nodes and blocks) was used to port the BCAS to cluster systems in the form of the PDCS emulator. The PDCS emulator is a static library using the message passing interface (MPI). The principle of its operation is to emulate the operation of the PDCS computational cores, using the communication network of a supercomputer for transferring tokens between cores [11].

III. SETS OF NODES AND BLOCKS OF THE MATCHING PROCESSOR

This section provides examples of sets of nodes and blocks of the matching processor, which change the structure of the computational core of the system and increase the execution efficiency of specific tasks.

A. Set of nodes and blocks for multi-input tokens

The tools that provide hardware support for the execution of multi-input nodes have been introduced into the command system of the matching processor of the PDCS (Fig. 2). The units of information circulating in the system are the token and the packet. The token belonging to multi-input nodes is called an "M-token", and the packet is called an "M-packet". "M-packet" in contrast to the standard packet (with two data fields) contains several data fields (depending on the size of the multi-input node).

For the effective use of "M-tokens" a corresponding set of nodes and blocks has been developed, which contains the node of "M-tokens" and "M-packets" processing, as well as new blocks in the CAMK and token memory. Increased efficiency is achieved due to the fact that only the key of the first incoming "M-token" is stored in the CAMKM (CAMK for "M-tokens"), which allows storing more keys of such tokens in the "expensive" content addressable memory. In the "cheap" token memory, implemented as a direct access memory, the data of the "M-tokens" are stored in accordance with their positions in the "M-packet". Each interaction changes the value of the token field "Data count of M-node". When all the "M-tokens" are collected, the data is extracted from the token memory and transferred for the "M-packet" generation.

This set of nodes and blocks was simulated by using the function of processing multi-input tokens. The operating time of the hardware nodes of "M-token" and "M-packet" processing is determined by the "M-token processing time" parameter.

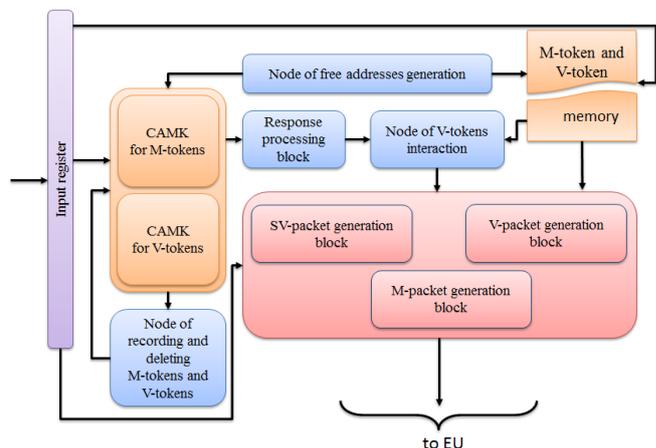


Fig. 2. Set of nodes and blocks of the matching processor for multi-input tokens and vector data (CAMK – content addressable memory of keys; EU – execution unit)

B. Set of nodes and blocks for vector data

In order to work with vectors, changes were made to the command system, as well as to the composition of the set of nodes and blocks of the matching processor (Fig. 2). Three new types of tokens, responsible for working with vector data, were introduced, and their hardware support was implemented in the MP.

New hardware elements related to the processing of vectors are introduced into the basic set of nodes and blocks: the unit of "V-token" processing, the unit of "V-token" and "V-packet" generation, the "V-tokens" memory.

The following tokens initiate the work of nodes and blocks: a vector ("V-token"), a new type of multi-input token ("M-token") and an activation token ("A-token"). New algorithms for the operation of the MP with "M-tokens" have been developed. If earlier N "M-tokens" formed the "M-packet", now one "V-token" is formed, which, without leaving the MP, comes on its input to search for paired "V-token" or "A-token". When two "V-tokens" interact, a "V-packet" is formed, containing $M_1 + M_2$ data, where M_1 and M_2 are the number of data of the first and second tokens, respectively. In the packet, the data that belongs to the token with the "left" interaction type are placed first, then the data that belongs to the token with the "right" interaction type.

The interaction reduces by one the multiplicity value of both tokens, except in the case when the multiplicity is infinite. The token whose multiplicity becomes 0 is deleted. The key of the remaining token (if its multiplicity is not exhausted) is written to the CAMK. When "V-token" and "A-token" interact, an "SV-packet" is formed, which is actually a multi-input packet that will contain as much data as there were in the "V-token" and will work with them as with separate elements of the vector.

This set of nodes and blocks was simulated by using the function of processing vector tokens. The operating time of the corresponding hardware nodes is determined by the "V-token processing time" parameter.

C. Set of nodes and blocks for special tokens

The task "Fast Fourier Transform" (FFT). The FFT algorithm is used in many application areas, including signal processing. The FFT task is a task with poor time localization and on traditional systems there are large delays

when performing memory access operations. This task is included in many test packages for verifying computing systems, and that is why studies have been carried out on the possibility of its effective implementation at the PDCS.

When solving this task on systems with a traditional architecture, between iterations of the algorithm it is necessary to perform barrier synchronization, which reduces the execution efficiency of the task. The PDCS through the use of the features of the computing model can simultaneously perform operations "butterfly" belonging to different iterations of the algorithm, which allows reducing standing time of the hardware and increasing the efficiency of using hardware resources.

In order to effectively implement this algorithm, a specialized set of nodes and blocks was developed (Fig. 3). The hardware node "FFT" was created, which performs the operation "butterfly", forms tokens for the next iterations, and also tracks the end of the program (by analyzing one of the token fields). A new type of token "FFT-token" has been added to the command system. In addition, a computation distribution function (a hash function) was created, which reduces the number of transfers between the computational cores, and its hardware support is added to the hash block.

The simulation of this nodes and blocks is performed by the function "Special operations". The operation time of each hardware FFT node is determined by the "FFT processing time" parameter.

Task "Multiplication of sparse matrices". The need for calculations using sparse matrices arises when solving optimization problems, when numerically solving partial differential equations, in graph theory and in many other vital scientific and engineering applications. When working with sparse matrices, one of the most pressing and complex tasks (from the point of view of implementation) is the operation of multiplying these matrices. The complexity of implementing an efficient matrix multiplication algorithm is due to the variability of the structure of their sparsity, which in turn affects the methods of storing and processing the values of matrix elements.

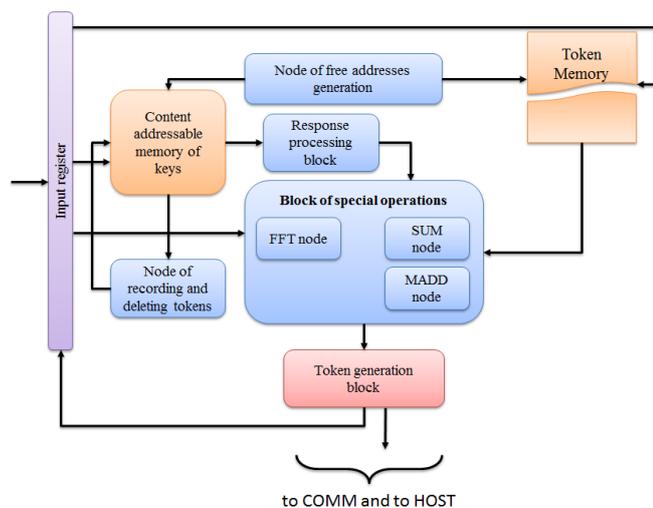


Fig. 3. Set of nodes and blocks of the matching processor for special tokens (FFT – "Fast Fourier Transform"; SUM – "Summation"; MADD – "Multiply-Accumulate"; COMM – commutator)

When designing a specialized system for the "Multiplication of sparse matrices" task on the basis of the PDCS, the multiplication and addition operations can be transferred from the execution unit to the MP. This is achieved through the use of special tokens, which at the hardware level replace entire chains of calculations.

For this, the hardware node "Summation" was implemented in the MP (Fig. 3). This node supports the operation of the special token "SUM-token". This allows the reduction of the program code by removing the program node that performs the summation of two products of matrix elements. It also saves time spent on creating a packet, sending it and subsequent processing on the execution unit.

Also, the hardware node "Multiply-Accumulate" and special token "MADD-token" were introduced into the MP. The hardware node initiates the mechanism of special tokens processing. The "MADD-token" performs the calculation of the product of two matrix elements, and the result is transmitted through the special "SUM-token" for the subsequent summation of all products. Such a mechanism allowed replacing the execution of the "matrix multiplication" dataflow program with hardware-software solutions. The simulation of this nodes and blocks is performed by the function "Special operations". The operation time of both hardware nodes is determined by the "Special-token processing time" parameter.

Introduction into the MP of such hardware nodes as "FFT", "Summation" and "Multiply-Accumulate", and the corresponding tokens into the command system allows completely abandon the execution unit in the composition of the PDCS computational core or reduce its composition to a simple device that will only form result tokens for returning them to the HOST machine.

IV. EXPERIMENTS

Studies of the sets of nodes and blocks, which are described in the previous section, were conducted on the behavioral cycle-accurate simulator (BCAS) and the emulator of the PDCS. The BCAS makes it possible to estimate the execution time (with an accuracy of one clock cycle) of the dataflow program on the selected architecture of the computing system. The specific architecture is created from the pre-created parameterized sets of nodes and blocks.

The test packet consisted of the following tasks: "Spatial Filter" (implementations based on standard two-input tokens and on multi-input tokens); "Composition Of Vectors" (using vector tokens and, accordingly, a set of nodes and blocks for vector data); "Fast Fourier Transform" (implementations using standard tokens and the special "FFT-token"); "Multiplication of sparse matrices" (implementations using standard tokens and special tokens "SUM-token" and "MADD-token").

Fig. 4 presents the results of tasks execution when using the developed sets of nodes and blocks. The highest speedup (compared to using a standard set of nodes and blocks, and, accordingly, standard tokens) is achieved using an FFT block. The speedup magnitude is 5.1 (the average magnitude using this set is 4.8). The use of multi-input tokens speeds up the execution time of these tasks by a maximum of 4.59 times (average magnitude is 4.3). A set of nodes and blocks for multiplying sparse matrices speeds up the task by a maximum of 2.39 times (the average magnitude is 1.93).

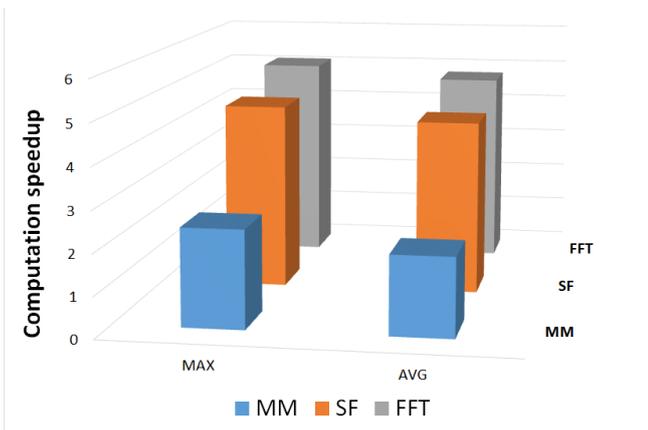


Fig. 4. Speedup of execution time when using the developed sets of nodes and blocks for the tasks "Multiplication of sparse matrices", "FFT" and "Spatial Filter" (MM - "Multiplication of sparse Matrices" task; SF - "Spatial Filter" task; FFT - "Fast Fourier Transform" task; MAX - maximum magnitude; AVG - average magnitude)

Experiments have shown that with the introduction of multi-input tokens and packets, the "Spatial Filter" task is solved much faster. Moreover, with an increase in the size of the image (compared to the two-input algorithm), this advantage grows.

For small vector dimensions, the execution time of the hardware implementation of the FFT task is reduced several times. With an increase in the dimension of the input vector, the gap between the execution time of hardware and software implementations only grows. At the same time, for a completely hardware implementation of the program, this speedup is nonlinear - on a small number of cores, the program is executed more than twice as fast.

V. CONCLUSION

When designing the PDCS architecture that implements the dataflow computing model, various variants of the sets of nodes and blocks of the matching processor were developed, which will allow the creation of computing systems for both specialized and universal applications.

The simulation of variants of sets of nodes and blocks of the matching processor of the PDCS computational core was carried out on a behavioral cycle-accurate simulator and the emulator using functions that described the operation of specific nodes within the matching processor.

The result of the conducted research is the confirmation by experiments of the correctness of the chosen direction of work aimed at optimizing the set of nodes and blocks of the matching processor of the PDCS computational core. The following can be chosen as optimization criteria: the task execution time, the area occupied by a set of nodes and blocks on a crystal, power consumption, the required amount of the content addressable memory, the versatility of using particular nodes and blocks. Experiments show that, by

changing the composition of the basic set of nodes and blocks, it is possible to improve the operation of the computational core for more efficient execution of tasks of various classes.

ACKNOWLEDGMENT

The reported study was funded by RFBR, project number 17-07-00478.

REFERENCES

- [1] A. D. Ivannikov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, D. N. Zmejev, "Dataflow Computing Model - Perspectives, Advantages and Implementation," in Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, Sept 29 - Oct 2, 2017, pp. 187-190.
- [2] Jack B. Dennis and David P. Misunas, "A preliminary architecture for a basic data-flow processor," in Proceedings of the 2nd annual symposium on Computer architecture (ISCA '75), ACM, New York, NY, USA, 1974, pp. 126-132. DOI=<http://dx.doi.org/10.1145/642089.642111>
- [3] A. P. W. Böhm, "Dataflow and hybrid dataflow architecture summary," in Parallel computer systems, Rebecca Koskela and Margaret Simmons (Eds.), ACM, New York, NY, USA, 1990, pp. 281-286. DOI=<http://dx.doi.org/10.1145/100215.100286>
- [4] B. Lee, A. R. Hurson, "Dataflow Architectures and Multithreading," Computer, Aug 1994, vol. 27, no. 8, pp. 27-39.
- [5] J. Silc, B. Robic, T. Ungerer, "Asynchrony in parallel computing: From dataflow to multithreading," Parallel and Distributed Computing Practices, 1998, vol. 1, no. 1, pp. 3-30.
- [6] Anup Das and Akash Kumar, "Dataflow-Based Mapping of Spiking Neural Networks on Neuromorphic Hardware," in Proceedings of the 2018 on Great Lakes Symposium on VLSI (GLSVLSI '18), ACM, New York, NY, USA, 2018, pp. 419-422.
- [7] Tony Nowatzki, Vinay Gangadhar, and Karthikeyan Sankaralingam, "Exploring the potential of heterogeneous von neumann/dataflow execution models", in Proceedings of the 42nd Annual International Symposium on Computer Architecture (ISCA '15), ACM, New York, NY, USA, 2015, pp. 298-310.
- [8] Xiaowei Shen, Xiaochun Ye, Xu Tan, Da Wang, Zhimin Zhang, Dongrui Fan, and Zhimin Tang, "POSTER: An Optimization of Dataflow Architectures for Scientific Applications," in Proceedings of the 2016 International Conference on Parallel Architectures and Compilation (PACT '16), ACM, New York, NY, USA, 2016, pp. 441-442.
- [9] T. Becker, O. Mencer, S. Weston, G. Gaydadjiev, "Maxeler Data-Flow in Computational Finance," in Proc. De Schryver C. (eds) FPGA Based Accelerators for Financial Applications, Springer, Cham, 2015.
- [10] A. D. Ivannikov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, D.N. Zmejev, "Global Distributed Associative Environment - Evolution of Parallel Dataflow Computing System "Buran", in Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2018), Kazan, Russia, Sept 14 - 17, 2018, pp. 655-659.
- [11] N. N. Levchenko, A. S. Okunev, D. N. Zmejev, A. V. Klimov, "Implementation of parallel dataflow computational model on cluster supercomputers," in Proceedings of the International IEEE EAST-WEST DESIGN & TEST SYMPOSIUM (EWDTS'2013), Rostov-on-Don, Russia, September 2013, pp. 394-396.

Optimized time-delayed feedback control of fractional chaotic oscillator with application to secure communications

Amir Rikhtegar Ghiasi

Faculty of electrical and computer engineering, University
of Tabriz, Tabriz, Iran
Email: agiasi@tabrizu.ac.ir

Mona Saber Gharamaleki

Faculty of electrical and computer engineering, University
of Tabriz, Tabriz, Iran
Email: m.saber96@ms.tabrizu.ac.ir

Elaheh Mohammadi asl Khasraghi

Faculty of electrical and computer engineering, University
of Tabriz, Tabriz, Iran
Email: elaheh.mhd96@ms.tabrizu.ac.ir

Zahra Sattarzadeh Kalajahi

Faculty of electrical and computer engineering, University
of Tabriz, Tabriz, Iran
Email: z.sattarzadeh96@ms.tabrizu.ac.ir

Abstract—In this paper new type of time-delayed feedback control is designed based on particle swarm optimization method. This controller is used in the control of chaotic behavior of fractional order chaotic oscillator. Using a discretization method practical realization of the proposed system has been done. Also secure communication using the designed method is done. Simulation results show the performance of the designed controller.

Keywords—Time-delayed feedback; Fractional-order; Chaotic oscillator; Particle swarm optimization; secure communications

I. INTRODUCTION

According to the chaotic behavior in many fields, study on this behavior is so important problem. Chaotic systems are unpredictable and this unpredictability has very application in many fields such as laser [1-2], biological systems[3] and secure communications [4]. Unpredictable behavior of these systems make control of these systems as a challenging problem. For the decades, studies on the control of chaotic systems have been done. In last decades, fractional version of chaotic systems have been expanded [5]. Fractional calculations are the general version of ordinary calculations, and according to the high accuracy on fractional order modeling of systems, this type of modeling has gained the great attractions. Especially on the chaotic systems, different types of fractional order systems are introduced [6-7].

Different types of controllers have been introduced according to this regard such as passive control [8-9], feedback control[10], impulsive control[11], and many others. The basic idea of these methods mostly obtained from two basic method: OGY method [12] and time delay feedback control [13].

Idea of time delay feedback control is add a control signal to the chaotic system to stabilize the system. This control signal is related to the present state and a delayed value of state. The main

work in this method is to determine proper feedback gain and time delay to guarantee the stability of the chaotic system[18].

A new method to stabilize the chaotic behavior of fractional order system has been introduced, recently [5].

Fractional order makes chaotic system with much complexity. And control of this system becomes a challenging problem. In this paper this chaotic behavior on fractional order system has been controlled using the optimized time-delayed feedback which is the new point in the chaos control in the fractional order systems.

The organization of this paper is as the follows: In section 2 basic relations about fractional calculations are introduced. Model of fractional order electrical chaotic system is shown in section 3. Section 4 describes the pervious studies on the time-delayed feedback control and optimized method. Practical implantation and simulation of the proposed introduced in section 5 and main conclusions of the paper are mentioned on section 6.

II. FRACTIONAL ORDER SYSTEMS

General forms of the fractional order operator have been defined in many studies [15]. One of the well known definitions is Riemann-Liouville definition as [16]:

$$D_t^\alpha f(t) = \frac{1}{\Gamma(n-\alpha)} \frac{d^n}{dt^n} \int_0^t \frac{f(\tau)}{(t-\tau)^{n-\alpha+1}} d\tau \quad (1)$$

where n is an integer such that $n - 1 \leq \alpha < n$.

Laplace transform of fractional derivate is defined as [16]:

$$L\{D_t^\alpha f(t)\} = s^\alpha L\{f(t)\} - \sum_{k=1}^{n-1} [D^{\alpha-1+k} f(t)]_{t=0} \quad (2)$$

where $L\{\}$ shows Laplace transform. For zero initial condition of fractional derivate of $f(t)$ Laplace transform of fractional derivative, operates as the integer one.

III. FRACTIONAL ORDER CHAOTIC OSCILLATOR

In this paper we consider an electrical chaotic oscillator (ECO) which was introduced on [4]. This system is written with the fractional order equations as the follows [19].

$$\begin{cases} D^\alpha x_1 = x_2 \\ \frac{dx_2}{dt} = x_3 \\ \frac{dx_3}{dt} = -a(x_1 + x_2 + x_3) + f(x_1) \end{cases} \quad (3)$$

which $\frac{d^\alpha}{dt^\alpha}$ shows the fractional derivative. $f(x_1) = \text{sign}(x_1)$ is the model nonlinearity. Block diagram of this system is shown on the Fig.1. System's behavior for $\alpha = 0.95$ and $a = 0.51$ is illustrated in the Fig.2.

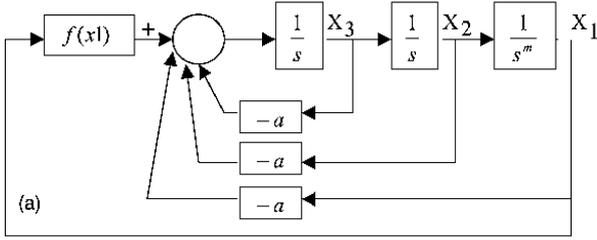


Fig. 1. Block diagram of fractional chaotic oscillator

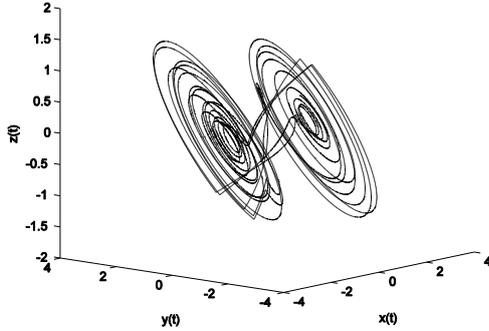


Fig. 2. Behavior of the fractional ECO

IV. OPTIMIZED TIME DELAYED FEEDBACK CONTROL

A. Time-delayed feedback control

In this part, first we introduced the basic method of time delayed feedback control. Then the optimized version of this method is proposed. We consider a n-th order fractional system as the follows:

$$\begin{cases} D_*^{\alpha_1} x_1(t) = f_1(x(t)) + U_1(t) \\ D_*^{\alpha_2} x_2(t) = f_2(x(t)) + U_2(t) \\ \vdots \\ D_*^{\alpha_n} x_n(t) = f_n(x(t)) + U_n(t) \end{cases} \quad (4)$$

where D_*^α shows the fractional derivative. x_i represents the states of system, and f is the vector which shows the nonlinear parts of each state of system and

$$U_i(t) = \sum_{j=1}^n K_{ij} [x_j(t-T) - x_j(t)] \quad (5)$$

is the delayed feedback force. This force is added to i th state of the system, K_{ij} are the feedback gains and T is the constant time delay [19].

System (4) without controller has a equilibrium point as $P = (x_1^*, x_2^*, \dots, x_n^*)$. The design of controller problem is to find appropriate gains and constant delay which makes the P as a locally asymptotically stable point using the forces as Eq.(5). Linearizing the system (4) around the P can be shown as:

$$\begin{pmatrix} D_*^{\alpha_1} \tilde{e}_1(t) \\ D_*^{\alpha_2} \tilde{e}_2(t) \\ \vdots \\ D_*^{\alpha_n} \tilde{e}_n(t) \end{pmatrix} = \hat{A} \cdot \begin{pmatrix} \tilde{e}_1(t) \\ \tilde{e}_2(t) \\ \vdots \\ \tilde{e}_n(t) \end{pmatrix} + \begin{pmatrix} \tilde{u}_1(t) \\ \tilde{u}_2(t) \\ \vdots \\ \tilde{u}_n(t) \end{pmatrix} \quad (6)$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \quad (7)$$

is the Jacobian, matrix at P , $\tilde{e}_i(t) = x_i(t) - x_i^*(t)$ shows the error from the equilibrium point. $\tilde{u}_i(t)$ shows the feedback control input which is defined as the follows.

$$\tilde{u}_i(t) = \sum_{j=1}^n K_{ij} [\tilde{x}_j(t-T) - \tilde{x}_j(t)] \quad (8)$$

Theorem 1.

The characteristic equation of system (3) is defined as:

$$\det[\Delta(s)] = 0 \quad (9)$$

where

$\Delta(s)$ is generated using the Laplace transform of Eq.(6).

The stability of origin shows the zero initial conditions for Laplace transform. If the following inequality is satisfied then system (3) is locally asymptotically stable around equilibrium point P

$$|\arg(s)| > \pi/2$$

where s shows the roots of characteristic equation.

Theorem 2. Without controller we consider system (3), A is the Jacobian matrix at P and P is an unstable equilibrium point. If the number of positive real eigenvalues of A was an odd number, then for time-delayed feedback control (4) cannot stabilize the unstable equilibrium with any controlling parameters.

B. Particle Swarm optimization

The main idea of this algorithm has been obtained from the behavior of animals such as birds. These swarms want to find the best location for food and when they find the food algorithm is finished. Each of solution in this algorithm called 'particle'. These particles have cost and the best particle is that has less cost. Then in each iteration worst particles are omitted and the better particles have updates. Particles have two characteristics, first one is the location and second one is velocity. In each iteration these characteristic for the remained particles have been updated using the following relations [14]:

$$\begin{aligned} V_{ij}(t+1) &= W * V_{ij}(t) + C_1 R_1 [p_j(t) - x_{ij}(t)] + C_2 R_2 [g_j(t) - x_{ij}(t)] \\ x_{ij}(t+1) &= x_{ij}(t) + V_{ij}(t+1) \end{aligned} \quad (10)$$

where V is the velocity of particle and x is the location of particle. $R1, R2$ are coefficients which in $[0,1]$ and $C1, C2$ are the learning rate. Flowchart of this algorithm is shown in the following.

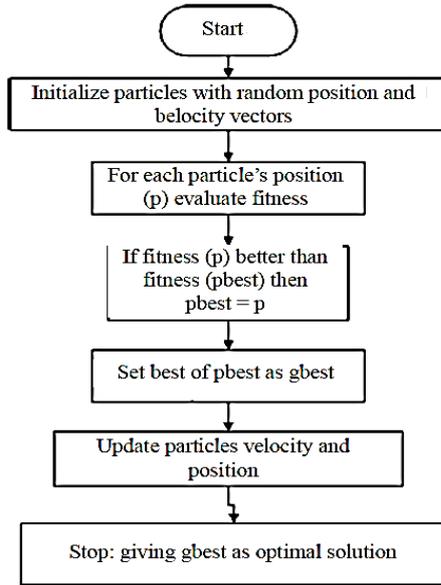


Fig. 3. Block diagram of PSO algorithm

C. Time-delayed feedback control based on particle swarm optimization

In this part for selecting the control parameters k_{ij} and T in the time-delayed feedback control we use PSO algorithm. This algorithm searches between the possible coefficients and select the best coefficients such that the following cost function becomes minimum:

$$f(t) = \int_0^T (\tilde{e}_1^2(t) + \tilde{e}_2^2(t) + \tilde{e}_3^2(t)) dt \quad (11)$$

where \tilde{e}_i for $i = 1,2,3$ are as the same as mentioned after Eq.(7).

D. Secure communication

In this part we use fractional order chaotic systems for the chaotic masking. Chaotic masking is one of the well-known algorithm in the information transmitting. Diagram of this method is shown in Fig.4. A chaotic system generates the carrier and this carrier combined with information signal and summation of two signals is transmitted through communication channel. In the receiver, chaotic synchronization is completed and after subtraction detected signal is obtained.

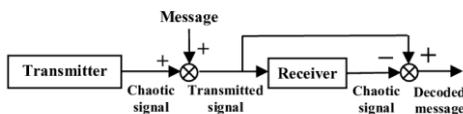


Fig. 4. Block diagram of secure communication process

V. SIMULATIONS AND PRACTICAL IMPLIMANTATIONS

Practical and simulation results of the proposed problem is investigated in this part. Practical implementation of fractional electrical chaotic oscillator is shown on the Fig.1. Practical implementation is done using an ARM microcontroller. Discretization of system (2) has been done using the proposed method in [16]. Phase portraits of practical implementation are shown in the Fig.6-8. After that the proposed TDFC has been shown in the Figs.9-11. Fig.8 shows the first time response of controlled system, Fig.10 shows the second state and Fig.11 shows the third states of system with TDFC method. The initial conditions for practical implementation is selected as $(x_1(0), x_2(0), x_3(0)) = (6.3, 12.5, 8)$.

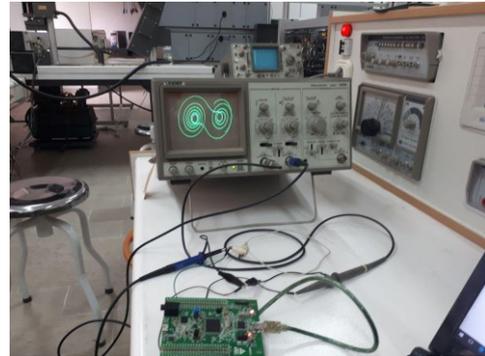


Fig. 5. Practical implementation of fractional ECO system

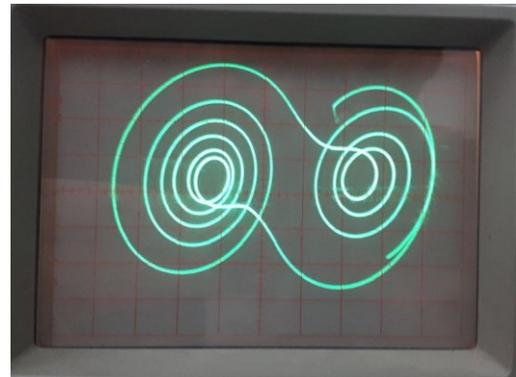


Fig. 6. Phase portirate of fractional ECO system $x_1(t) - x_2(t)$



Fig. 7. Phase portirate of fractional ECO system $x_1(t) - x_3(t)$



Fig. 8. Phase portirate of fractional ECO system $x_2(t) - x_3(t)$

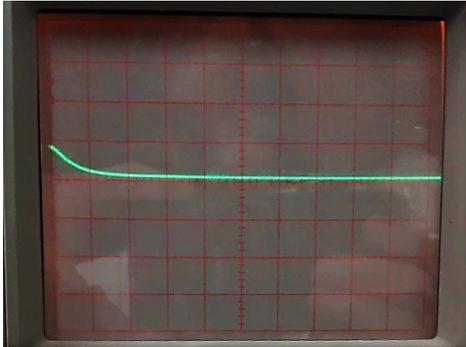


Fig. 9. Practical view of the first state of system(4) using TDFC

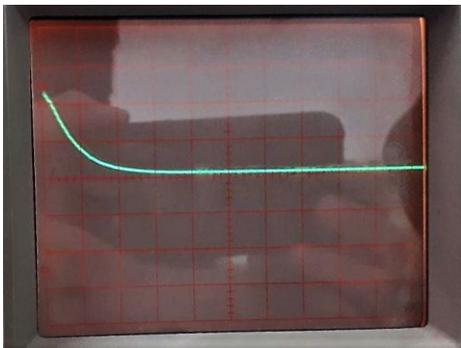


Fig. 10. Practical view of the second state of system(4) using TDFC

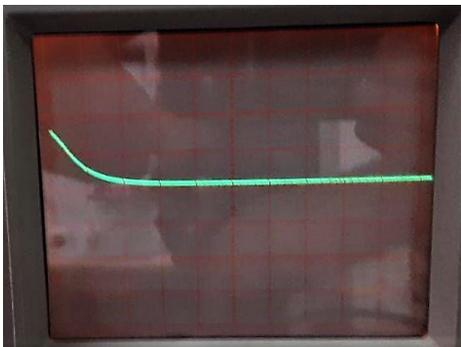


Fig. 11. Practical view of the third state of system(4) using TDFC

For comparison we simulated the proposed method with the general TDFC method. Initial conditions for system (2) in simulation is selected as (0.1, 0.2, 0.3). Number of iterations for PSO algorithm is selected 50 iterations. Convergence of PSO algorithm is shown in Fig.12. Results of simulation of

simple and optimized TDFC method for the system (2) are illustrated in Fig.13-15.

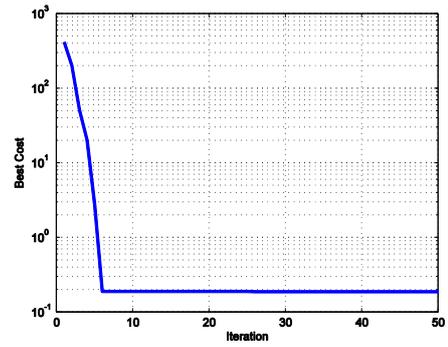


Fig. 12. Convergence of PSO algorithm

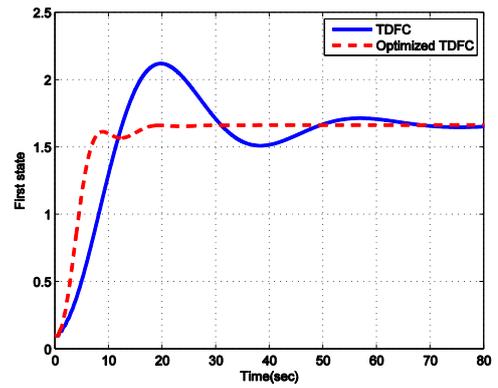


Fig. 13. First state of controlled fractional ECO system

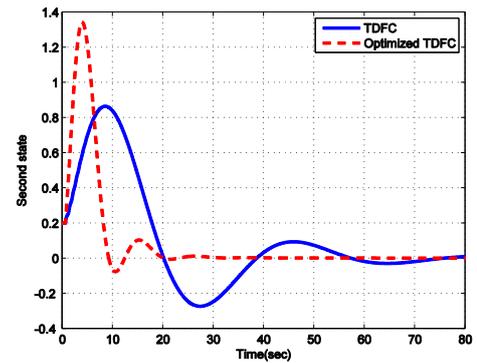


Fig. 14. Second state of controlled fractional ECO system

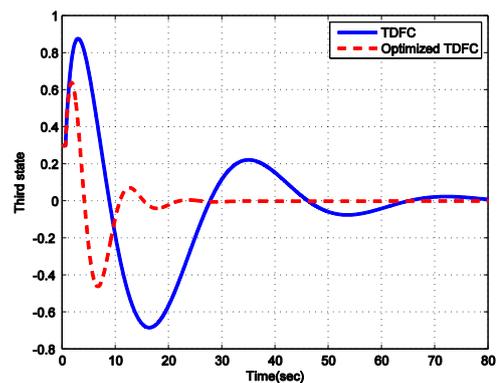


Fig. 15. Third state of controlled fractional ECO system

In the next step of simulation, secure communication problem with two fractional order chaotic system is investigated. Three signals are considered as $F(t) = [\sin(10t) + 5 \cos(3t); 7 \cos(4t) + \sin(8t); 4 \cos(3t) + \sin(t)]$ Based on proposed method three signals are transferred with chaotic masking and those signals are received at receiver as seen in the Fig.16-18.

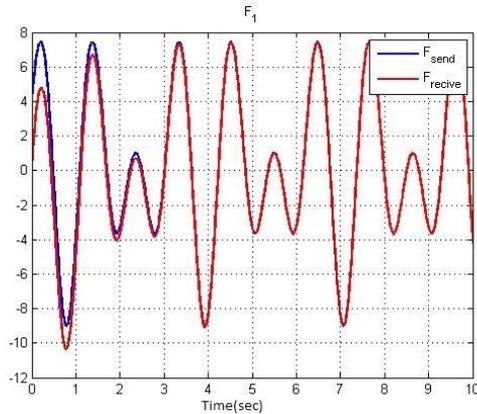


Fig. 16. Secure communicate of first signal

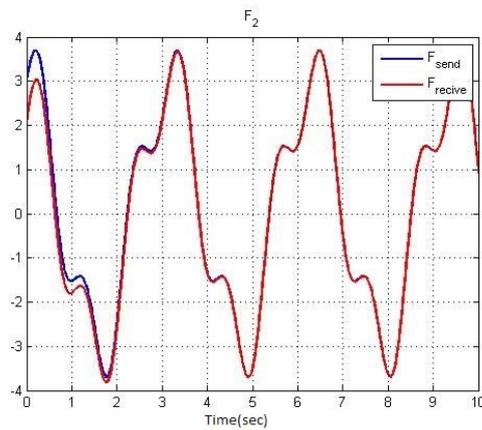


Fig. 17. Secure communicate of second signal

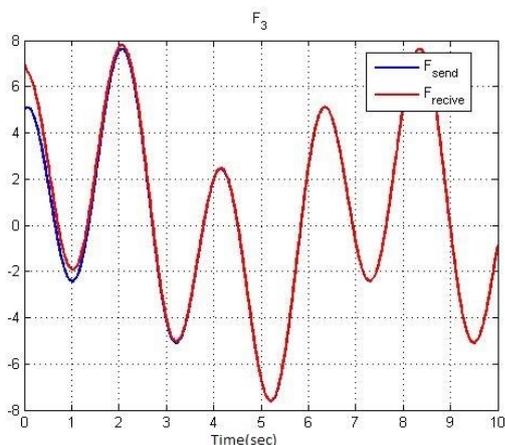


Fig. 18. Secure communicate of third signal

VI. CONCLUSION

In this paper practical realization of fractional order chaotic system was investigated. New optimized time-delayed feedback control method was used for the stabilizing the chaotic behavior of fractional order electrical oscillator. Optimizing method was based on the particle swarm optimization. Practical and simulation results proved the performance of the proposed method. Finally secure communication using the proposed method has been done. Results showed the performance of the designed method.

REFERENCES

- [1] DeShazer DJ, Breban R, Ott E, Roy R. Detecting phase synchronization in a chaotic laser array. *Physical Review Letters*. 2001 Jul 6;87(4):044101.
- [2] Roy R, Murphy Jr TW, Maier TD, Gills Z, Hunt ER. Dynamical control of a chaotic laser: Experimental stabilization of a globally coupled system. *Physical Review Letters*. 1992 Mar 2;68(9):1259..
- [3] Shekofteh Y, Jafari S, Sprott JC, Golpayegani SM, Almasganj F. A gaussian mixture model based cost function for parameter estimation of chaotic biological systems. *Communications in Nonlinear Science and Numerical Simulation*. 2015 Feb 1;20(2):469-81.K.M. Cuomo , A.V. Oppenheim , *Phys. Rev. Lett.* 71 (1993) 65 .
- [4] Yang T. A survey of chaotic secure communication systems. *International journal of computational cognition*. 2004 Jun;2(2):81-130.
- [5] Behinfaraz R, Badamchizadeh MA. New approach to synchronization of two different fractional-order chaotic systems. *International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 2015 Mar 3 (pp. 149-153).
- [6] Behinfaraz R, Badamchizadeh M. Optimal synchronization of two different in-commensurate fractional-order chaotic systems with fractional cost function. *Complexity*. 2016 Sep;21(S1):401-16.
- [7] Behinfaraz R, Ghaemi S, Khanmohammadi S. Adaptive synchronization of new fractional-order chaotic systems with fractional adaption laws based on risk analysis. *Mathematical Methods in the Applied Sciences*. 2019 Apr;42(6):1772-85..
- [8] Mahmoud GM , Mahmoud EE , Arafa AA . Passive control of n-dimensional chaotic complex nonlinear systems. *J Vib Control* 2013;19(7):1061-71 .
- [9] Behinfaraz R, Ghiasi AR. A survey on reliability analysis in controller design. *14th International Colloquium Signal Processing & Its Applications (CSPA)*, 2018 on 2018 Mar 9 (pp. 198-202).
- [10] Yang T , Chua LO . Impulsive stabilization for control and synchronization of chaotic systems: theory and application to secure communication. *IEEE Trans Circuits Syst I* 1997;44(10):976-88 .
- [11] Risk assessment in control of fractional-order coronary artery system in the presence of external disturbance with different proposed controllers
- [12] Pyragas K . Continuous control of chaos by self-controlling feedback. *Phys Lett A* 1992;170(6):421-8 .
- [13] Kazemi, Ali, Reza Behinfaraz, and Amir Rikhtegar Ghiasi. "Accurate model reduction of large scale systems using adaptive multi-objective particle swarm optimization algorithm." *Mechanical, System and Control Engineering (ICMSC)*, 2017 International Conference on. IEEE, 2017.
- [14] Behinfaraz, Reza, and Mohammad Ali Badamchizadeh. "Synchronization of different fractional-ordered chaotic systems using optimized active control." *Modeling, Simulation, and Applied Optimization (ICMSAO)*, 2015 6th International Conference on. IEEE, 2015.
- [15] Behinfaraz, Reza, Mohammadali Badamchizadeh, and Amir Rikhtegar Ghiasi. "An adaptive method to parameter identification and synchronization of fractional-order chaotic systems with parameter uncertainty." *Applied Mathematical Modelling* 40.7-8 (2016): 4468-4479.
- [16] Pyragas K. Delayed feedback control of chaos. *Philos Trans R Soc A* 2006;364:2309-34.
- [17] Mahmoud, Gamal M., et al. "Chaos control of integer and fractional orders of chaotic Burke-Shaw system using time delayed feedback control." *Chaos, Solitons & Fractals* 104 (2017): 680-692.

[18] Gjurchinovski A, Sandev T, Urumov V. Delayed feedback control of fractional-order chaotic systems. *Journal of Physics A: Mathematical and Theoretical*. 2010 Oct 13;43(44):445102.

[19] Behinfaraz R, Badamchizadeh MA. Synchronization of different fractional order chaotic systems with time-varying parameter and orders. *ISA transactions*. 2018 Jul 24.

Implementation Variants of the Global Distributed Associative Computing Environment for the Parallel Dataflow Computing System “Buran”

Nikolay Levchenko
*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
nick@ippm.ru

Anatoly Okunev
*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
oku@ippm.ru

Dmitry Zmejjev
*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
zmejjev@ippm.ru

Abstract—For the high-performance parallel dataflow computing system (PDCS) “Buran” this article proposes to move away from dividing the computational space into basic computational cores and moving on to organizing the system from associative computational cores. The use of the global distributed associative computing environment changes the architecture of the system and creates significant advantages compared with the basic PDCS architecture. The article discusses the implementation variants of the matching processor and the entire PDCS architecture in the form of a global distributed associative environment of information processing. The article describes the testing of the distributed associative environment in the behavioral cycle-accurate simulator, which was performed on various tasks. In the course of the experiments, the behavior of the new computing system architecture with hardware failures was also studied.

Keywords—*global associative computing environment, matching processor, associative computational core, matching of keys, fault tolerance*

I. INTRODUCTION

In the 1980s, work was carried out on the architectures of computing systems that implement the dataflow computing model [1]. Such a model was conceived to be closer to the hardware in comparison with the von Neumann computing model. The developers claimed that it uses the “natural” parallelism inherent in hardware much better [2-4]. Nevertheless, problems in technological production and the untimeliness of this approach, which was much ahead of its time, led to the actual closure of all works in this area [5-6].

In recent years, developers of supercomputers, realizing the advantages of the dataflow computing model, resumed research of non-traditional computing models for their use in extreme-scale computing [7-11]. This is confirmed by ongoing conferences, new articles that are devoted both to solving hardware problems and calls dictated by the software part.

The parallel dataflow computing system (PDCS) “Buran” implements the dataflow computing model with a dynamically formed context. The PDCS overcomes many problems in the development and use of supercomputers, the main of which is low actual performance when scaling tasks up to millions of computational cores.

Along with solving the problem of paralleling computations while creating high-performance computing

systems, it is required to create a set of measures to ensure reliability in the computation process. In order to improve the basic architecture of the PDCS, the authors propose to move away from the currently used approach of dividing the computational space in the PDCS into the basic computational cores and move to the concept of organizing the entire computing system from the so-called associative computational cores. Where, one associative computational core must be physically located in a single processor (chip). The new architecture of the PDCS “Buran” was called the architecture of the global distributed associative computing environment.

The goal of the work is to consider the implementation variants of the matching processor of the computational core and the entire PDCS architecture in the form of the global distributed associative computing environment of information processing.

II. ARCHITECTURE OF THE DISTRIBUTED ASSOCIATIVE COMPUTING ENVIRONMENT

The computational core of the basic PDCS architecture contains an execution unit, a matching processor, a packet and token buffer, a hash block, and an internal commutator [12].

One of the key elements of the computational core is the matching processor (MP). The MP includes various blocks: content addressable memory of keys, descriptor memory, token memory, hash block for distributing computations by stages, and others.

The MP compares the keys of the task tokens [12], ensuring the implementation of the basic principles of the dataflow computing model and preventing overflow or underload of hardware resources in the computation process.

The main functions of the matching processor:

- organization of computational processes;
- activation of computations by data readiness;
- synchronization of computational processes by data;
- organization of group selection of operands by mask;
- management of computing resources using special operations (in particular, hardware support for control commands);

- control of the state of hardware resources and the rate of computational processes.

The content addressable memory of keys (CAMK) is intended for storing keys with masks, fixing recorded keys in the vacant register, generating addresses of free cells for storing keys, and also finding matches with the keys stored in it.

The use of the CAMK makes it possible to extract from the task the entire parallelism (even not discovered by the programmer), that exists in the algorithm for solving this task [13]. This allows the efficient utilization of the available hardware resources of the PDCS “Buran” and the achievement of a high degree of actual tasks scaling on systems consisting of hundreds of thousands of computational cores. Also, the CAMK actually performs multiple comparison operations and condition checks.

The negative characteristic of CAMK is high energy consumption. However, a lot of research is currently being conducted aimed at solving this problem - the creation and use of energy-efficient CAM [14-17].

A feature of the CAMK used in the PDCS is the presence of both external and internal masks. The mask allows the comparison of token keys only in the area that is not closed by it (not masked). The masked area of the key is ignored and is considered as matched, regardless of the content. The advantage of the mask applied to the token key is that by using it the multiple responses are organized (the key matched the keys of more than two tokens at the same time). This mechanism leads to significant memory savings and a reduction in the number of tokens generated (the load on data transmission lines is reduced), since, for example, to multiply a vector by a number, it is necessary to send not M numbers (where M – the size of a vector), but only one number (token) with the corresponding mask.

All the described functions of the MP are preserved in the new architecture of the GDACE.

A. New variant of the PDCS “Buran” architecture.

The use of the global distributed associative computing environment, in fact, changes the PDCS architecture, and these changes, which lie within the framework of the dataflow computing model, provide significant advantages compared to the basic PDCS architecture.

The architecture of the global distributed associative computing environment (GDACE) is a set of associative computational cores (ACC), connected by a token commutator. Each ACC consists of a set of local associative computational elements (LACE), each of which performs the function of a computational core in the basic PDCS architecture. Such a structure allows a programmer who creates tasks in the dataflow programming paradigm to take into account only the ACC numbering when setting up or choosing distribution functions.

Fig. 1 presents the implementation variant of the GDACE architecture [18].

The features of the GDACE architecture are the following:

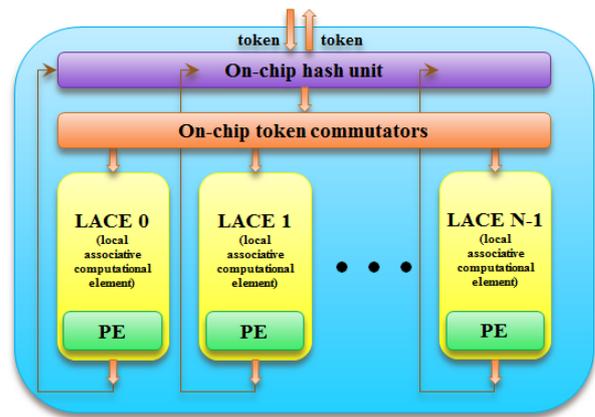


Fig. 1. The architecture of the GDACE associative computational core

- in contrast to the basic PDCS architecture, the distribution of computations in the GDACE architecture is carried out up to an individual processor (chip);
- the GDACE architecture provides more convenient mechanisms for organizing a heterogeneous architecture (the heterogeneity of the processors may consist both in a different compositions of the processor itself, and in a different number of computational cores) of the computing system when developing hardware and creating software;
- the GDACE architecture increases the reliability of the computing system due to the possibility of its operation when individual LACEs fail;
- the GDACE architecture allows the work with ACC containing LACEs, the number of which may be not a multiple of 2.

So, the advantages of the GDACE are increased fault tolerance, manufacturability, architectural flexibility, increased performance and simpler parallel programming.

III. VARIANTS OF ASSOCIATIVE ENVIRONMENT IMPLEMENTATION ON THE PROGRAM MODEL

A parameterized implementation of the GDACE architecture was created on the behavioral cycle-accurate simulator of the PDCS [19]. With the use of the developed program model of the GDACE architecture, a comparison was made with the basic PDCS architecture.

During the model-based analysis several variants for the GDACE implementation were considered.

The first implementation of the GDACE is the division of a single associative space of token storage into a set of specialized blocks, each of which corresponds to a specific type of tokens that do not interact with each other at the hardware level.

In the behavioral cycle-accurate simulator, separate blocks were created for standard tokens without a mask, a block for storing tokens with a mask, a block for storing multi-input tokens and a block for vector tokens.

This implementation of the associative environment allows the process of different types of tokens in parallel and to ensure maximum performance of the matching processor when using different types of tokens.

The second implementation of the GDACE is the extension of the first one. It consists in expanding the matching processor by a set of functional devices. The introduction of new types of tokens and functional devices that provide hardware support for the processing of tokens with built-in computing functions at the level of content addressable memory makes it possible to increase the efficiency of the computing system on a number of simple operations without the need to form packets and then process them in execution units.

The third implementation of the GDACE differs from the first two in the possibility of parallel and pipeline processing of input tokens. The computational core can receive tokens from the communication network, from the execution unit of their ACC, and, in addition, a new token can be formed directly in the matching processor of the LACE.

In this variant of the GDACE implementation a buffer CAMK is added at the entry of tokens to control the interaction between incoming tokens. The CAMK is replaced with a pipelined CAMK for five types of tokens: standard without a mask, multi-input, vector, global and universal. Universal token is a standard type of token with a mask. The response processing scheme and the token and packet generation unit are implemented for each direction.

The variant of the pipeline GDACE allows the division of token matching process into stages. That significantly saves energy, reducing the number of “false” matchings (such matchings that do not result in the formation of a packet) between the individual bits of tokens keys.

This means that at the same time at the same stage of the pipeline operation can be three different tokens when the following conditions are met:

- the received tokens must be of different types, that is, each of them is compared in its CAMK;
- parallel processing of tokens does not violate the logic of the command system of the associative environment;
- no busy signal in the direction to the relevant CAMK.

In parallel, without additional conditions the following tokens can be processed: standard without a mask, multi-input and vector. For parallel processing of other tokens, additional control of interaction with each other is required.

IV. EXPERIMENTS

In the course of the research, parametrizable program modules of three variants of the GDACE implementation were created.

Testing of the distributed associative environment on the program model was carried out on such tasks as “Matrix Multiplication”, “Molecular Dynamics” [20] and “Spatial Filter”.

The first implementation of the GDACE architecture was tested on the “Spatial Filter” task. Variants of the task with standard and multi-input tokens were executed. Multi-input tokens were processed in a separate block of the CAMK.

Experiments have shown a fivefold reduction in the execution time of the task on the input data of the same size (356,920 cycles against 77,665) on 16 computational cores. The amount required for the work of the CAMK also

decreased 4 times due to the use of multi-input tokens (in this case, one key in the CAMK accounts for all the tokens in the multi-input packet).

The second implementation of the GDACE was tested on the “Molecular Dynamics” task. For this task, a new type of token “MD-token” was introduced, intended for hardware filtering of “empty” pairwise interactions. The functional device responsible for processing this type of tokens is also implemented.

The results showed that with the use of this variant of the GDACE implementation, the time to complete the task is reduced by 10%.

The third variant of the GDACE implementation was tested on the “Matrix Multiplication” task. Parallel processing of a large token flow from different destinations made it possible to reduce the processor matching time by 5 times (compared to the basic PDCS architecture).

During the experiments, the behavior of the new computing system architecture with hardware failures was studied on the “Matrix Multiplication” and “Molecular Dynamics” tasks (the computing system configuration – 4 processor modules containing four computational cores in each module). In particular, the failure of a single computational core in the processor was simulated. The task of multiplying matrices with the dimension of 64 * 64 elements in the variant without failures was executed in 493,161 cycles, and in the variant where the failure of the computational core was simulated – 554,569 cycles for the basic PDCS architecture and 493,502 cycles for the GDACE when a single ACC fails. For the “Molecular Dynamics” task, the results are as follows: 1,304,506 cycles on the system without failures, 1,966,168 cycles for the basic PDCS architecture when a single computational core fails and 1,744,357 cycles for the GDACE when a single ACC fails (Fig. 2).

It can be stated that the proposed architecture of the distributed associative computing environment copes better with the failure of individual computational cores and the performance loss is less on the GDACE than on the basic architecture of the PDCS. With the increase in the number of processors, the effect becomes even more noticeable.

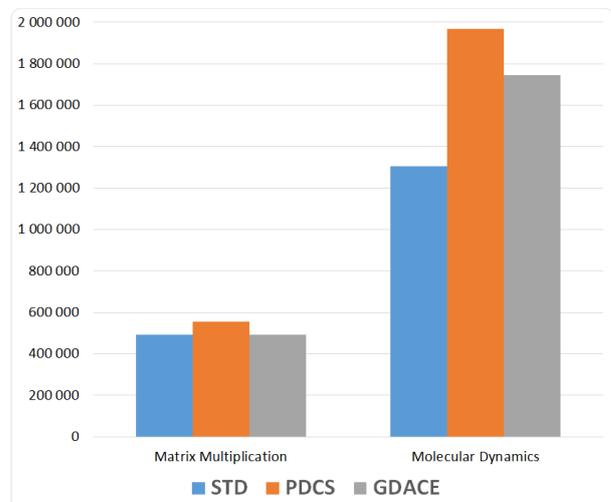


Fig. 2. Comparison of the execution time of the tasks "Molecular Dynamics" and "Matrix Multiplication" in case of failure of a single computational core in the PDCS and GDACE (STD – basic PDCS architecture)

V. CONCLUSION

At present, the dataflow computing model is actually experiencing its rebirth, since the traditional computing model encounters certain difficulties in its use for parallel computation.

In the GDACE architecture, that implements such model, the distribution of computations within each processor can be carried out in a special way (independent of other computational cores). This allows the computing system to function even with heterogeneous processors. If in the process of a computing system operation, associative computational cores and even their individual elements (execution units, memory, etc.) fail, the process of hardware failure will not affect the distribution of computations. Or more precisely, there is no need to change distribution function in the event of a failure. This is due to the fact that in the event of a failure, a redistribution of computations within the processor module at the hardware level will be automatically made.

The GDACE architecture is mainly applicable in the following cases:

- when a high reliability of computations is required;
- when using heterogeneous processor modules with different qualitative and quantitative composition of nodes and blocks;
- in case of insufficient yield of suitable crystals in the manufacture process due to lower requirements for processor functionality.

It is also worth noting that, at present, the field of artificial intelligence and the associated machine learning tasks are acquiring special relevance. In the tasks of machine learning a special place is occupied by neural networks and their deep learning. It is expected that just in this direction, the dataflow computing model has certain advantages, especially in parallelizing computations with sparse data, since matrix-vector multiplication is actively used when learning neural networks.

REFERENCES

[1] Jack B. Dennis and David P. Misunas, "A preliminary architecture for a basic data-flow processor," in Proceedings of the 2nd annual symposium on Computer architecture (ISCA '75), ACM, New York, NY, USA, 1974, pp. 126-132. DOI=<http://dx.doi.org/10.1145/642089.642111>

[2] A. P. W. Böhm, "Dataflow and hybrid dataflow architecture summary," in Parallel computer systems, Rebecca Koskela and Margaret Simmons (Eds.), ACM, New York, NY, USA, 1990, pp. 281-286. DOI=<http://dx.doi.org/10.1145/100215.100286>

[3] B. Lee, A. R. Hurson, "Issues in Dataflow Computing," *Advances in computers*, 1993, vol. 37, pp. 285-333.

[4] R. S. Nikhil, G. M. Papadopoulos, and Arvind, "T: a multithreaded massively parallel architecture," in Proceedings of the 19th annual international symposium on Computer architecture (ISCA '92), ACM, New York, NY, USA, 1992, pp. 156-167. DOI=<https://doi.org/10.1145/139669.139715>

[5] B. Lee, A. R. Hurson, "Dataflow Architectures and Multithreading," *Computer*, Aug 1994, vol. 27, no. 8, pp. 27-39.

[6] J. Silc, B. Robic, T. Ungerer, "Asynchrony in parallel computing: From dataflow to multithreading," *Parallel and Distributed Computing Practices*, 1998, vol. 1, no. 1, pp. 3-30.

[7] J. Fresno, D. Barba, A. Gonzalez-Escribano, D. R. Llanos, "HitFlow: A Dataflow Programming Model for Hybrid Distributed- and Shared-

Memory Systems," *Int J Parallel Prog*, 2019, vol. 47, issue 1, pp. 3-23. DOI=<https://doi.org/10.1007/s10766-018-0561-2>

[8] Hadi Alizadeh Ara, Amir Behrouzian, Martijn Hendriks, Marc Geilen, Dip Goswami, and Twan Basten, "Scalable Analysis for Multi-Scale Dataflow Models," *ACM Trans. Embed. Comput. Syst.*, Aug. 2018, vol. 17, issue 4, article 80. DOI=<https://doi.org/10.1145/3233183>

[9] Jani Boutellier, Henri Lunnikivi, "Design Flow for Portable Dataflow Programming of Heterogeneous Platforms," in 2018 Conference on Design and Architectures for Signal and Image Processing (DASIP), 2018. DOI=<https://doi.org/10.1109/DASIP.2018.8596931>

[10] M. Solinas, "The TERAFLUX Project: Exploiting the DataFlow Paradigm in Next Generation Teradevices," in 2013 Euromicro Conference on Digital System Design, Los Alamitos, CA, 2013, pp. 272-279. DOI=<https://doi.org/10.1109/DSD.2013.39>

[11] Tony Nowatzki, Vinay Gangadhar, Newsha Ardalani, and Karthikeyan Sankaralingam, "Stream-Dataflow Acceleration," in Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA '17), ACM, New York, USA, 2017, pp. 416-429. DOI=<https://doi.org/10.1145/3079856.3080255>

[12] A. D. Ivannikov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, D. N. Zmejev, "Dataflow Computing Model – Perspectives, Advantages and Implementation," in Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, Sept 29 - Oct 2, 2017, pp. 187-190.

[13] N. N. Levchenko, A. S. Okunev, D. N. Zmejev, "Solutions to Problem of CAM Overflow in the Parallel Dataflow Computing System "Buran",," in Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2018), Kazan, Russia, Sept 14 - 17, 2018, pp. 649-654.

[14] Catherine E. Graves, Wen Ma, Xia Sheng, Brent Buchanan, Le Zheng, Si-Ty Lam, Xuema Li, Sai Rahul Chalalasetti, Lennie Kiyama, Martin Foltin, John Paul Strachan, and Matthew P. Hardy, "Regular Expression Matching with Memristor TCAMs for Network Security," in Proceedings of the 14th IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH '18), ACM, New York, NY, USA, 2018, pp. 65-71. DOI=<https://doi.org/10.1145/3232195.3232201>

[15] Yue Zha and Jing Li, "Liquid Silicon: A Data-Centric Reconfigurable Architecture Enabled by RRAM Technology," in Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '18), ACM, New York, NY, USA, 2018, pp. 51-60. DOI=<https://doi.org/10.1145/3174243.3174244>

[16] Amanda F. Fonseca, Douglas L. Willian, Thiago R. B. S. Soares, Luiz G. C. Melo, and Omar P. Vilela Neto, "CAM/TCAM - NML: (ternary) content addressable memory implemented with nanomagnetic logic," in Proceedings of the 30th Symposium on Integrated Circuits and Systems Design: Chip on the Sands (SBCCI'17), ACM, New York, NY, USA, 2017, pp. 174-179. DOI=<https://doi.org/10.1145/3109984.3110004>

[17] Rekha Govindaraj and Swaroop Ghosh, "Design and Analysis of STTRAM-Based Ternary Content Addressable Memory Cell," in *J. Emerg. Technol. Comput. Syst.*, May 2017, vol. 13, issue 4, article 52., DOI=<https://doi.org/10.1145/3060578>

[18] A. D. Ivannikov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, D.N. Zmejev, "Global Distributed Associative Environment - Evolution of Parallel Dataflow Computing System "Buran",," in Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2018), Kazan, Russia, Sept 14 - 17, 2018, pp. 655-659.

[19] N. N. Levchenko, A. S. Okunev, D. N. Zmejev, "Development Tools for High-Performance Computing Systems Using Associative Environment for Computing Process Organization," in bk.: Proceedings of IEEE EAST-WEST DESIGN & TEST SYMPOSIUM (EWDTS'2016), Yerevan, Armenia, October 14-17, 2016, pp. 359-362.

[20] Shrinidhi Hudli, Shrihari Hudli, Raghu Hudli, Yashonath Subramanian, and T. S. Mohan, "GPGPU-based parallel computation: application to molecular dynamics problems," in Proceedings of the Fourth Annual ACM Bangalore Conference (COMPUTE '11), ACM, New York, NY, USA, 2011, article 10, 8 pages. DOI=<https://doi.org/10.1145/1980422.1980432>

Researching Resilience a Holistic Approach

Zoya Dyka, Ievgen Kabin and Peter Langendörfer
IHP – Leibniz-Institut für innovative Mikroelektronik
Im Technologiepark 25
Frankfurt (Oder), Germany

Abstract— With the advent of the Internet of things and 5G the number of devices that are controlling parts of our lives increases dramatically. As we rely on these devices it is essential that they work properly over long times and under unpredictable conditions. If a system can ensure the aforementioned properties the system is considered to be resilient. In this paper we discuss the idea of a holistic approach that covers redundancy, reliability and security of individual components up to complex systems to networked cyber physical systems of systems. We also introduce the preliminary work done at IHP on which we build our resilience approach.

Keywords— resilience, resilience engineering, cyber physical system (CPS), e-health.

I. BACKGROUND AND MOTIVATION

IT Systems are penetrating our everyday lives with ever increasing speed. The reason for this is that people want to have an easy life empowered by many more or less pervasive systems. On the one hand this trend is making our lives more comfortable on the other hand we rely on those systems and if they collapse all the comfort and maybe even essential services are no longer available. The issue is that while scaling technologies allow devices to become even smaller and more energy efficient the number of faults increases. In addition, those devices are connected with each other to fulfil more complex tasks. It is predicted that the number of these devices will reach 50 billion already in 2020 [1]. The application areas cover:

- Smart home
- Industry 4.0
- Agriculture 4.0
- Medicine 4.0
- Autonomous driving
- etc.

As the networked devices are there to assist human beings it is expected that these services are always available. It is essential that the devices and the services they provide are under full control and cannot cause any harm. This needs to hold true for unintended faults as well as for intended faults or attacks.

The rest of this paper is structured as follows. The next section discusses the requirements a resilient system needs to fulfil as well as the approaches that are needed to ensure such behaviour. Section 3 introduces our idea of a holistic approach toward resilience. The pre-existing building blocks we intend to use for our example application area i.e. e-health are presented in section 4. The paper closes with conclusions in section 5.

II. RESILIENCE REQUIREMENTS

Given the laziness of mankind and the application fields it is of utmost importance that these systems are providing their services independent of the operating conditions. As the latter are changing over time these systems need to be empowered to react on these changes to keep at least a certain level of those services. So the vision is that the devices used are working:

- forever
- smooth
- without interaction /intervention of the end users
- and are low cost.

In other words we expect those systems to provide extreme high dependability or at least fail safe operation under adverse conditions. In addition these systems need to be tamper resistant to protect their users' privacy.

But in reality these systems face different types of disturbances, faults and outages. So during the system development engineers need to anticipate these issues and to implement appropriate remedies to ensure proper service qualities at least up to a certain extent once above mentioned issues appear. The standard means is redundancy which is or at least has been very successful in the past.

The vast majority of the devices is based on silicon chips. Semiconductors are sensitive to changes in their environment e.g. temperature, humidity, heat dissipation, electromagnetic emanation. Changes in these parameters are normal and the devices can cope with these changes in a certain range. But the sensitivity of the devices increases with scaling technologies which means that problems caused by the environmental parameters will increase more often and their prediction will become more complex. Of course the safety margins may be increased so that faults induced by the environmental parameters will be avoided or at least their occurrence can be reduced. This normally comes with increased cost which may be tolerated in some application areas such as e-health but might be not in others. The other effect that may occur is that the performance may be reduced, which may cause issues in real time applications such as automation control.

It is common knowledge that faults can and will appear even in case of significant robustness is built in the system. So, in order to provide service in presence of faulty devices or part of devices they need to be handled at run time. This means the devices need to have monitoring and repair capabilities. Devices that cannot recover after entering a faulty state need to be

removed from the system of systems. So the cyber physical system of systems (CPSoS) need self-x features such as:

- self-diagnosis;
- self-repair;
- self-re-configuration;
- etc.

Beneath the hardware induced faults of single devices many issues are caused by the communication if the devices are part of a network. To remedy these issues devices may be removed or added to the network. In the best case these networks are self-organizing.

One of the main challenges is privacy. The basic properties for ensuring privacy are confidentiality and data integrity. They can be achieved by applying cryptographic means. The cryptomeans used today are proven to be secure from the mathematical point of view. But this is based on the fact that the keys are kept secret. This becomes really tricky if the current through the cryptographic device or the electromagnetic emanation of the device while executing the cryptographic operations can be measured. There are plenty well known side channel attacks. Especially dangerous are passive attacks as they may go undetected by the victims. This means the victim still assumes his/her system to be secure. The problem is that the victim may lose control fully, or just loses his/her data and/or know how. For system engineers this means they need – in addition to faults – to take these attacks into account during the development phase. This task is extremely demanding as it requires detailed knowledge and experience in different domains of expertise for example materials, technology processes, cryptographic algorithms, communication protocols. But nevertheless full protection is almost infeasible.

TABLE I. shows challenges of cyber physical systems of systems as well as potential solutions. Here we focus on self-x features as they empower to the systems to cope with upcoming challenges by themselves. This is of crucial importance as many issues will appear unexpectedly.

TABLE I. POTENTIAL CAUSES OF FAULTS/PERFORMANCE ISSUES AND COUNTERMEASURES.

Problem-Causes	Solutions
Energy	low-power
Environmental and working parameters natural and/or intelligent fluctuations: t°, heat dissipation, light, EM-Pulses, radiation, aging, ...	redundancy self-monitoring self-aware self-calibrating self-(re)configuration self-(re)optimizing self-adaptive self-healing
System(s) joining /leaving of components	self-managing self-organizing self-protection self-upgrading self-modifying ...
Wireless communication	
Information leakage passive observations/attacks	

III. THE RESILIENCE APPROACH

When talking about resilience it is important that any type of recovery functionality needs to be provided at run time. This holds especially true for the self-x properties that we mentioned earlier. The final goal is to cope with disturbances in a fully transparent manner. What we mean by this is that even though faults cannot be avoided their impact shall be fully compensated, i.e. the end user will not notice that there is any challenge for/in the system. In order to achieve such transparency the CPSoS needs to be capable to predict at least some of the disturbances and to react appropriately even before the issues really occur. Such a CPSoS would be resilient.

In order to enable a system to become resilient means on all layers are required. By layers we do not mean protocol layers but components of a CPSoS from material, via manufacturing technologies and analogue/digital design up to software and network engineering. Fig. 1 shows using IHP's structure and activities how such a vertically aligned resilience support could look. In addition on the left hand side of Fig. 1 the idea of a specific resilience design methodology is shown. We consider such a methodology essential as the design of resilient systems faces new currently unanticipated challenges. The especially demanding issue is to predict potential issues and to react on these. Even more important is that with the term resilience we mean that the system can react on challenging situations on its own. The essentially new/innovative aspect here is that when features such as fault tolerance are engineered all potential reactions of the system are predefined. When it comes to resilience engineering the challenge is that the system needs to react not according to prescribed recipes but finding its own strategy to adapt to adverse conditions, hopefully making all challenges transparent for the end user, i.e. keeping the service level at an acceptable level.

The core feature needed to achieve the here mentioned resilience feature is *self-awareness* i.e. the knowledge about the current system state, faults, energy available, ongoing attacks, tasks to be fulfilled etc. While gathering the relevant information is even though challenging quite straight forward, including appropriate "sensors" such as hardware performance counters, environmental sensors etc. The real challenge is to develop a model that gives the sensor data at a certain point in time a semantics so that sensor data is transformed into information. This information is then the basis for any further decision.

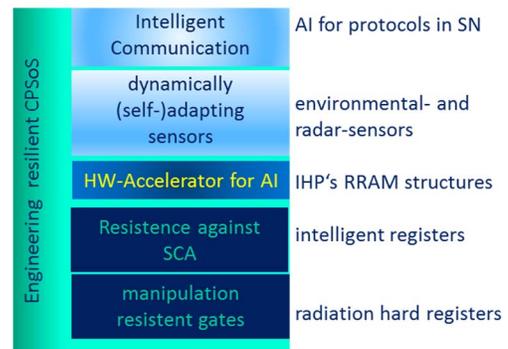


Fig. 1. Vertical approach inside IHP indicating technologies developed that can be used to build a resilient system.

We are aware of the fact that the idea we just explained is extremely complex and for sure there is *no one-size-fits-all* solution. So, we are focussing on a specific application area, e-health in our case. We are convinced that the design principles we are investigating using this example can be generalized for other application areas.

IV. RESEARCH VEHICLE E-HEALTH SYSTEMS

We decided to use e-health as the first application area in which we want to investigate resilience. This is due to the following facts, e-health systems ask for:

- Extremely high fault tolerance to ensure reliable service at all times and under all also adverse conditions;
- Extremely high security to ensure data confidentiality, data integrity and finally the patients' privacy.

In addition the example is pretty intuitive for everyone not only for skilled scientists. The other more practical reason is that we already implemented several e-health devices at IHP so that we have a reasonably solid fundament on which to build new research.

In the myAirCoach project [2] we used our own sensor node Ghost in which we integrated hardware accelerators for elliptic curve cryptography, AES and SHA-1. So in principle this sensor node can support confidentiality and authorization. We are still improving the accelerator with respect to tamper resistance. Ghost supports Bluetooth Low Energy for networking, see Fig. 2.

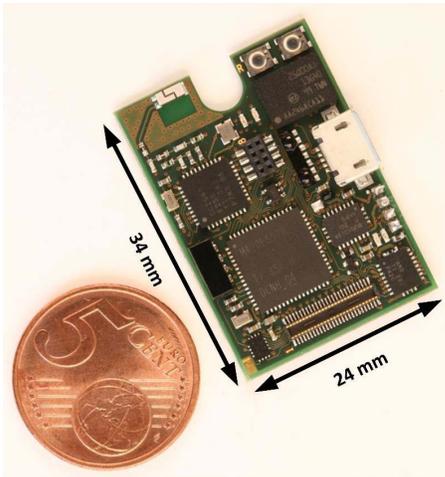


Fig. 2. Ghost sensor node developed in IHP supporting cryptographic means such as ECC, AES and SHA-1 (copied from [3]).

IHP developed a sensor to determine the viscosity of saliva for assessing the patients' health state concerning Chronic Obstructive Pulmonary Disease (COPD) [4], see Fig. 3 for parts of the device (a) and the fully assembled on (b). In order to ensure longevity the sensor needs self-calibrating features to ensure correct measurement results even after longer use.

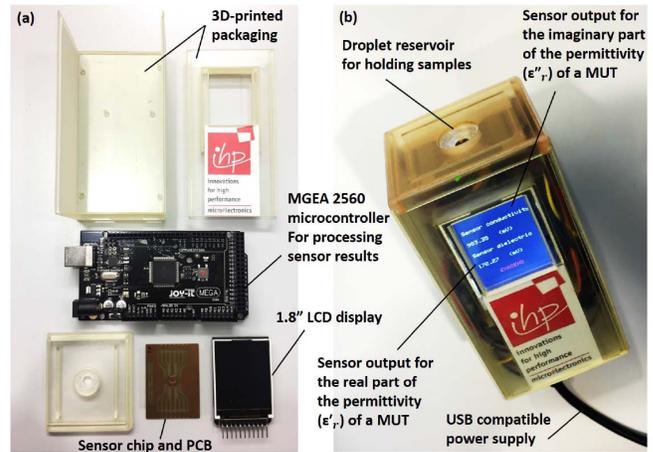


Fig. 3. Sensor for measuring permittivity of saliva for indicating COPD (a) parts of the device and (b) the fully assembled device (copied from [4]).

In addition to that IHP works towards radiation hard gates [5] that can be used to improve fault and manipulation resistance of ASICs.

So some of the essentially needed building blocks are available. In addition we gathered first experience with artificial intelligence means. Neural networks were used to classify the results of COPD sensors [6] while we evaluated means such as k-means for their appropriateness for side channel attacks [7].

But simply combining these building blocks does not make a system resilient. There are sophisticated means needed that help to select the most appropriate building blocks for a certain application. In addition means to assess the level of resilience are needed.

V. CONCLUSIONS

In this paper we introduced the idea to develop a holistic approach for making cyber physical systems of systems resilient. This idea is driven by the fact that resilience is a complex beast in the sense that the system has to cope with unpredicted situations while it consists of more or less reliable components that may fail at any time. We consider IHP extremely well suited to dare this endeavour as IHP has already researched reliability and security from components up to complex systems. These building blocks have also been introduced in this paper. We consider the orchestration of these blocks based on the also to develop self-awareness as one of the major challenge for upcoming years.

REFERENCES

- [1] Amy Nordrum, Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated, 18 Aug 2016, <https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated>, last viewed Mai 2019.
- [2] The myAirCoach Project: Analysis, modelling and sensing of both physiological and environmental factors for the customized and predictive self-management of Asthma, <http://myaircoach.eu/>, last viewed Mai 2019
- [3] Steffen Ortman, "Tragbare Sensoren zur Erfolgsbewertung in der Schlaganfallstherapie, FP-7 project StrokeBack," 18th LEIBNIZ CONFERENCE OF ADVANCED SCIENCE – SENSORSYSTEME 2014, Lichtenwalde, October, 2014, Germany, <https://leibniz-institut.de/Konferenzen/Sensorsysteme-2014/LK18-Ortman.pdf>

- [4] P. Soltani Zarrin, F. Ibne Jamal, N. Roeckendorf, and C. Wenger, "Development of a Portable Dielectric Biosensor for Rapid Detection of Viscosity Variations and Its In Vitro Evaluations Using Saliva Samples of COPD Patients and Healthy Control," *Healthcare (Basel)*, vol. 7, no. 1, Jan. 2019.
- [5] R. Sorge, J. Schmidt, C. Wipf, F. Reimer, R. Pliquet, and T. Mausolf, "J1CG MOS transistors for reduction of radiation effects in CMOS electronics," in *2018 IEEE Topical Workshop on Internet of Space (TWIOS)*, 2018, pp. 17–19.
- [6] P. Soltani Zarrin, C. Wenger, "Pattern Recognition for COPD Diagnostics Using an Artificial Neural Network and Its Potential Integration on Hardware-based Neuromorphic Platforms," submitted to 28th International Conference on Artificial Neural Networks (ICANN 2019), 17th – 19th September, 2019, Munich, Germany.
- [7] I. Kabin, M. Aftowicz, Y. Varabei, D. Klann, Z. Dyka and P. Langendörfer, "Horizontal Attacks Using k-means: Comparison with Traditional Analysis Methods," accepted for 10th IFIP International Conference on New Technologies, Mobility & Security (NTMS 2019), June 2019. Canary Island – Spain.

Modeling and debugging tools development for recurrent architecture*

Dmitry Khilko
Institute of Informatics Problems
Federal Research Center
"Computer Science and Control"
of the Russian Academy of
Sciences
Moscow, Russia
DHilko@yandex.ru

Yury Stepchenkov
Institute of Informatics Problems
Federal Research Center
"Computer Science and Control"
of the Russian Academy of
Sciences
Moscow, Russia
YStepchenkov@ipiran.ru

Yury Shikunov
Institute of Informatics Problems
Federal Research Center
"Computer Science and Control"
of the Russian Academy of
Sciences
Moscow, Russia
YIshikunov@gmail.com

George Orlov
Institute of Informatics Problems
Federal Research Center
"Computer Science and Control"
of the Russian Academy of
Sciences
Moscow, Russia
Orlov.jaja@gmail.com

Abstract—An unconventional multi-core recurrent data-flow architecture, that is being developed at Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRS CSC RAS) was successfully tested on digital signal processing domain both at the model level and on a hardware sample. Based on the test results, several mechanisms had been identified that required improvement and a decision was made to investigate the architecture on other subject domains. Software and main architectural blocks debugging are carried out with the specially developed hardware and software modeling tools. The active extension and debugging of the architecture by using these tools revealed a number of shortcomings of the existing software. To eliminate these shortcomings, two problems have to be solved: to provide a high degree of reconfigurability of the architecture's imitational model (to debug its mechanisms) and implement a symbolic modeling mode (to debug its software). The redesigning results of modeling and debugging tools for recurrent data-flow architecture are discussed in the article.

Keywords—*data-flow, modeling, debugging, recurrent architecture*

I. INTRODUCTION

In Russia, the team at the Department of Architecture and Circuit Design of Innovative Computing Systems of FRS CSC RAS is developing multi-core recurrent data-flow architecture (MRDA) [1]. The architecture of the MRDA is radically different in its main points not only from the classical architecture of von Neumann but also from other non-conventional architectures. Such architectures are characterized by the presence of two flows: instruction flow (active flow for von Neumann architecture) and data flow [2, 3] (active flow for data-flow architecture). The developed architecture is characterized by *self-sustained* data – a data representation structure where data and instruction flows are combined into one indivisible self-sufficient data flow. This feature provides an increase in the MRDA processing speed by reducing the number of steps required to execute the instruction. The "instruction fetch" step is implicit, and is executed simultaneously with the data processing.

Conventional data-flow architectures [2, 3] were developed to support applications with massive parallelism and dynamic

*The study was done by a grant from the Russian Science Foundation (Project №. 19-11-00334)

task distribution. At the same time, the token length (number of service bits) greatly exceeds the length of informational bits, which is unacceptable for economic reasons in the DSP domain.

The goal of MRDA development was to transfer the principles of data-flow architectures into the domain of DSP, taking into account its principal features and the requirements of economic efficiency. For most DSP applications, dynamic allocation is redundant [4]: the predictability of program execution time makes statistical allocation methods viable. The degree of parallelism of the main DSP applications is limited and ranges from 4 to 16 [5]. That is confirmed by the modern DSPs of Texas Instruments [6], with offerings ranging from four to eight-core configurations.

The implementation of these features and the introduction of special mechanisms for the uniform data compression into the MRDA made it possible to drastically reduce the token length while retaining the main principle of data-flow architectures — data readiness processing. Moreover, the recurrent data-flow computational model has a high potential for reducing redundancy due to its main feature: *recurrency*. It is the way for computational process organization when its progress depends on the current state and "unfolds" during the computation.

While developing architecture's prototype, it became necessary to create specialized hardware and software tools for modeling and debugging both the basic functional units and software. A set of hardware and software tools HARSF IDE [1] has been created as a result. Using these tools, we have managed to develop and debug imitational and hardware (VHDL) models of the architecture, as well as provide a fairly complete software development life cycle.

In order to test the architecture, a demonstrational task – the recognition of isolated word-commands has been chosen. As shown in [7], the results of architecture approbation both on software and hardware model confirmed the high potential of its effectiveness for the selected task. Therefore, the hardware sample based on the VHDL-model of the architecture has been developed in FPGA technology using the Cyclone V GT Development Kit.

The following abstractions (entities) were introduced into the imitational model architecture:

- Stage (stage of the conveyor) – implements the design patterns "Facade" and "Abstract Factory", provides an interface for accessing and controlling the main elements of the conveyor;
- Component – implements the "Facade" and "Abstract Factory" design patterns, provides an interface for accessing and controlling the internal structure of components;
- Block – the basic structural element of the model, implements a specific basic architecture mechanism, interacts with other blocks through the ports;
- Port – the main mechanism for data exchanging process organization between structural elements of the architecture.

This approach to the imitational model organization allows to make functionality changes quickly and to reconfigure the architecture to test new mechanisms. It is easy to see that the specific version of the architecture configuration is a graph with two types of vertices: "Port" and "Block", while the abstractions "Component" and "Stage" are subgraphs. Thus, it has become possible to use the "Graph" package (used to build various kinds of graphs) in a different context, which indicates a successful refactoring of HARSP IDE.

As can be seen from the description of the "ISIMPRABlock" interface each block of a model can have a theoretically unlimited number of both input and output ports. Therefore, a universal interface and a universal algorithm of block operation were developed. Each block must have at least three input ports: L-input, R-input, and Configuration-input, - as well as one output port: Result-output. In case R-input is unnecessary it should be set to "NON". Fig. 2 shows the universal block interface.

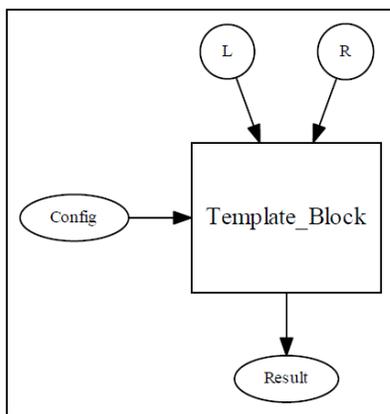


Fig. 2. Universal "Block" interface

Thus, the configuration of the model is reduced to the creation of a graph of blocks and ports. Fig. 3 demonstrates the "Juggler" component configuration. For greater clarity, numbers are replacing some of the port names.

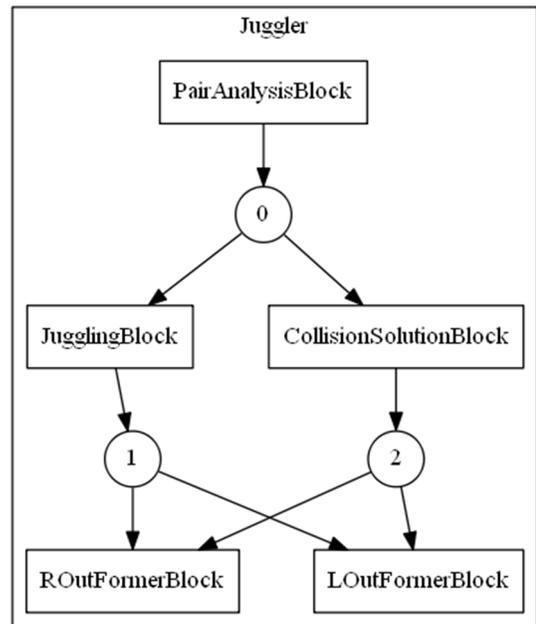


Fig. 3. "Juggler" component configuration

Being represented as a graph, the structure of the imitational model is much more consistent with the structure of the hardware model. This, in turn, means that the redesigning of the hardware model based on the imitational model is greatly simplified. Thus, the refactoring of imitational modeling tools resulted in a toolkit creation, which significantly simplified not only the development and debugging of the basic mechanisms of MRDA, but also the changes migration to the hardware implementation.

IV. SYMBOLIC MODELING

Another important task of the development of the modeling tool is the implementation of the symbolic modeling mode. This mode is necessary for debugging specialized software. Despite the fact that symbolic programming is natural for data-flow architectures, we have not been implementing this functionality for a long time. This was due to the sufficient power of the numerical simulation and debugging mode. However, with the increasing complexity of software and approbation of new subject domains, the need in a symbolic mode has aroused.

A. Symbolic data-flow graph

The main data structure, that we use for the implementation of symbolic modeling mode is a symbolic data-flow graph. This graph contains information about the computational process progress. In addition, a symbolic data-flow graph is the main tool for verification of software intended for MRDA.

A key feature of this graph is a fairly accurate visual interpretation of the imitational model's block structure. In order to achieve a high level of consistency between the visual elements of the data-flow graph and the architecture blocks, a special context notation for describing the graph cells has been developed. Fig. 4 demonstrates the context structure.

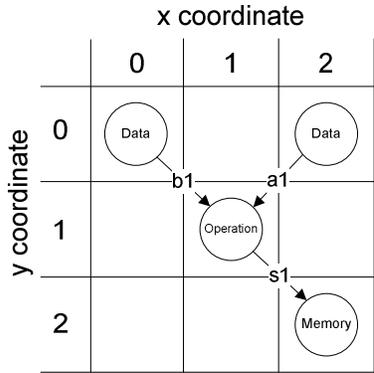


Fig. 4. Data-flow graph cell context

Computational units of MRDA have a superscalar architecture. Therefore, the context of one modeling step can contain: up to three data sources, up to three operations, and single memory write. Each of these elements must be displayed in the context of a single computational step at a data-flow graph. Therefore, to fulfill the layout requirements for all the elements of the data-flow graph, it was decided to form the context of a single step of calculations for one section in the form of a 3x3 grid, where each cell can contain a node.

In accordance with the requirements, three node types were introduced:

- "data" – represents the input data;
- "operation" – represents an operation performed in the current context;
- "memory" – represents the name of data, stored at internal memory blocks.

Data-flow graph arcs display the movement of data within and between contexts. Therefore, arc names are associated with the data they transfer. Using such a format for describing graph cells allowed us to enter and process the most complete amount of symbolic information about the computational process progress. In addition, a special XML representation was developed that can be easily serialized to a file and deserialized into a data structure into memory for later use in the modeling process.

B. Symbolic data representation

In order to implement symbolic modeling, a special data representation has been developed. Currently, the hardware sample of architecture supports 16-bit fixed-point data. Therefore, a "FixPoint16bit" library has been created for the imitational model, which implements 16-bit fixed-point mathematics. The FixPoint16bit class wraps the inner data object, that implements the "InnerData" interface. While inner data object wraps data with the symbolic name and other useful information. Fig. 5 shows a diagram of base classes and interfaces of this library.

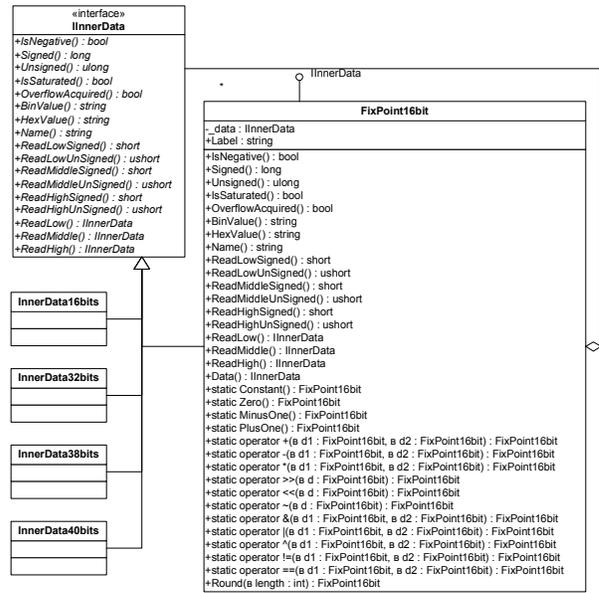


Fig. 5. "FixPoint16bit" package diagram

The "Name" property of the "InnerData" interface allows accessing the symbolic data name. In turn, the symbolic name can be obtained from three sources: the input data name, the arc name of the data-flow graph, or generated according to some rules. This library is designed in such a way that it can be easily replenished with classes representing a higher capacity data (32, 64 bits) or even with a floating point. This allows quick and easy creation of MRDA configurations and debugging them.

C. User feedback

To assess the results of the symbolic modeling mode implementation, we have collected feedback from software developers for MRDA. We were interested in the following questions in comparison with the previous version:

- How has the workflow changed?
- How has the iteration time requirements changed?

Toolset users within our department have been interviewed, and the following reviews have been received:

- The majority of users rated the new version as significantly more convenient due to the visualization of the programming process in the accustomed symbolic form.
- Some of the users noted the improvement of the iteration cycle speed.

V. CONCLUSION

As part of the development of the MRDA modeling tools, the following results were obtained:

- block architecture of the imitational model has been designed, providing wide options for reconfiguration and functionality extension;

- symbolic modeling mode has been implemented, that allows debugging specialized software efficiently.

The updated version of HARSP IDE imitational modeling subsystem will be used to redesign a number of the architecture mechanisms, as well as to implement the set of algorithms included at BTDImark2000 DSP benchmark. This benchmark will allow evaluating the effectiveness of the recurrent architecture in the digital signal processing subject domain.

ACKNOWLEDGMENT

In conclusion, we want to acknowledge Diachenko Y.G. and Morozov N.V. for invaluable contribution in development of MRDA.

REFERENCES

- [1] Yu. Stepchenkov, D. Khilko, Yu. Diachenko, Yu. Shikunov and D. Shikunov. Software and hardware testing of data-flow recurrent digital signal processor // Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2016), Yerevan, October, 14 - 17, 2016. pp. 168-171.
- [2] Arvind, Nikhil R.S. Executing a program on the MIT tagged-token data-flow architecture // IEEE Trans. Computer. – 1990. – V. 39. - №3 – P. 300 – 318.
- [3] D. N. Zmejev, A. V. Klimov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, Dataflow computing model as a paradigm of future mainstream of software development // Informatika i ee primeniya – informatics and applications, 2015, vol 9, issue 4, pp. 29–36. (In Russian).
- [4] K.Kronlof, O.Simula and J.Skytta. DFSP: A Data flow Signal Processor. IEEE Transactions on Computers, vol. C-35, issue 1, January 1986, p. 23-33.
- [5] Yu. V. Rogdestvenski, Yu. G. Diachenko. Fundamental parallelization estimates in voice signal processing systems. Sistemi i sredstva informatiki. – M: Nauka, 2002. – Issue. 12. – pp. 250 – 254. (In Russian).
- [6] TI DSP. Available at: <http://www.ti.com/processors/digital-signal-processors/products.html> (accessed 30.07.2019).
- [7] Khilko, Yu. Stepchenkov, D. Shikunov, Yu. Shikunov. Recurrent data-flow architecture: technical aspects of implementation and modeling results // Problems of Perspective Micro- and Nanoelectronic Systems Development - 2016. Proceedings / edited by A. Stempkovsky, Moscow, IPPM RAS, 2017. Part II. pp. 59-64.
- [8] Yu. Stepchenkov, Yu. Shikunov, N. Morozov, G. Orlov, D. Khilko. Hybrid Multi-Core Recurrent Architecture Approbation on FPGA // Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 2019 IEEE. P. 1705 – 1708.
- [9] BDTI DSP Kernel Benchmarks. Available at: <https://www.bdti.com/services/bdti-dsp-kernel-benchmarks> (accessed 30.07.2019).
- [10] Erich Gamma, Richard Helm, Ralph Johnson, John M. Vlissides. Stock Image. Design Patterns: Elements of Reusable Object-Oriented Software // Published by Addison-Wesley Professional, - 1994. 368 P. (ISBN 10: 0201633612).

Caution: GALS-ification as a Means against SCA Attacks

Zoya Dyka, Ievgen Kabin, Dan Klann, Frank Vater and Peter Langendoerfer
IHP – Leibniz-Institut für innovative Mikroelektronik
Im Technologiepark 25
Frankfurt (Oder), Germany

Abstract— In this paper we revisit the idea of asynchronism as a means to improve the resistance of hardware implementations of cryptographic algorithms. Here we are focusing on GALS designs. We use such a GALS implementation based on our synchronous design realized by colleagues for fine grained investigations. Our key finding is that a straight forward mapping of blocks of a synchronous design into GALS islands is not improving the designs resistance against SCA. This is due to the fact that the vulnerabilities of the initial design survive the transformation into the GALS design and make it vulnerable as well.

Keywords— ECC, side channel analysis (SCA) attack, horizontal attacks, simple power analysis (SPA), GALS.

I. INTRODUCTION

Elliptic Curve Cryptography (ECC) can guarantee not only confidentiality of communication but can also be used for authentication of persons/devices. Nowadays ECC is worldwide implemented for use in the Internet of Things, WSN, automation industry, protection of critical infrastructures, etc.

The core operation by ECC is the elliptic curve point multiplication with a scalar, denoted as kP operation. P is an EC point and k is a long binary number – the scalar. It is the private key in EC authentication protocols or a random number in the ECDSA signature generation protocol [1]. The kP operation is a time consuming operation and is often implemented in hardware with the goal to accelerate the computations. The algorithm for (mutual) authentication that is mostly implemented in hardware is the Montgomery kP algorithm using Lopez-Dahab coordinates [2] for EC over binary extended fields. The Montgomery kP algorithm is a bitwise processing of the scalar k . The scalar k is the secret that attackers try to reveal, often called the “key”.

The cryptographic strengths of a cipher algorithm may depend according to the definition of Kerckhoff only on the used key that is kept secret [3]. This means a potential attacker may know the algorithm itself, the plain text, the encrypted text and even the length of the key. In such a situation the attacker can test different numbers in order to reveal the key. Such an attempt is called brute force attack and the attacker needs to test 2^n numbers in the worst case to get a key of length n . The average number of attempts is 2^{n-1} . The cryptographic keys have to be sufficiently long so that the time the attacker needs for brute

forcing (or any other more efficient attack if known) is long. In this case the cryptographic approaches using long keys are assumed to be secure.

The situation changes dramatically if the attacker knows not only the input and output values but also intermediate values of the executed cryptographic operation. The intermediate values depend on the processed key. If an attacker can observe the intermediate values clockwise, it simplifies the key extraction significantly. If the attacker gets physical access to the device running the cipher algorithm, the intermediate data or even the key can be measured directly on chip using special equipment, for example by microprobing. The attacker can obtain knowledge about intermediate values indirectly, measuring physical parameters influenced/affected by the working chip e.g. the execution time of the analysed cryptographic operation, the clockwise energy consumption and its distribution during the execution of the operation, temperature, electromagnetic emission etc. The physically measurable parameters are a kind of “side effects”. Because all these parameters depend on the given input and processed key, these “side effects” can be analysed with the goal to reveal the key. Mostly analysed are current through the cryptographic chip while the kP operation is performed or the electromagnetic emanation of the chip, i.e. attackers can measure and analyse Power Trace(s) or Electromagnetic Trace(s) (EMT) of kP execution(s). Designers can do the analysis at the earlier stage of the designing using simulated PTs, i.e. designers attack the chips with the goal to evaluate their resistance against SCA attacks. Especially dangerous are so-called single trace attacks: Simple Power Analysis (SPA) attacks, horizontal differential analysis attacks for example [4], Horizontal Collision Correlation Analysis (HCCA) attacks [5]. By SPA the different operation sequences for the processing of a ‘0’ and ‘1’ bit value of the key cause different power profiles in the PT. The differences can be seen with eyes only, i.e. the processing of a key bit value ‘0’ is distinguishable from the processing a key bit value ‘1’. The balancing, regularity, atomicity principle as well different kinds of randomizations can be applied to avoid the key revealing, at least to reduce the success of attacks.

SCA attacks are a significant threat for implementations of cryptographic algorithms. Usually the cryptographic designs are implemented as synchronous circuits. Attackers can often concentrate on some operation(s) in the main loop of the Montgomery kP algorithm. Selected operations are performed

usually periodically if no special countermeasures are implemented, for example a randomization of the execution sequence. Implementing the Montgomery kP algorithm as an asynchronous design can increase its resistance against SCA significantly but the area and the energy consumption of a such implementation will be enormous. The split between these two strategies – synchronous and asynchronous design – can be the Globally Asynchronous Locally Synchronous (GALS) architecture. This strategy was proposed in [6] to countermeasure SCA attacks, especially SPA attacks.

In this paper we show that the GALS-ification strategy can be but is not necessarily a means to reduce the success of attacks. In section II we describe shortly the synchronous IHP kP accelerator for the EC B-233 and demonstrate how easy an SPA attack, with minimal knowledge about the implemented algorithm can be performed. In section III we show the separation of the synchronous design for a straight forward GALS-ification. In section IV we demonstrate the successful scalar revealing running an SPA against the simulated PTs of the GALS-ified kP design. The paper finishes with short conclusions.

II. SPA OF THE SYNCHRONOUS kP DESIGN

In 1999 Lopez and Dahab published in [2] an efficient kP algorithm that is based on the Montgomery observation [7] using special Lopez-Dahab projective coordinates of EC points. All calculations are performed using only the x -coordinate of the points. The most complex operation, i.e. the division of elements of binary Galois fields, was avoided. Only at the end of the kP algorithm the division of the field elements is necessary to recover the affine y -coordinate of the output point. These optimizations reduce the execution time and the energy consumption of the kP calculation significantly. Due to these advantages the Montgomery kP algorithm for EC point multiplication using Lopez-Dahab coordinates is a world-wide used algorithm for EC over extended binary fields $GF(2^l)$, especially in hardware implementations. Algorithm 1 shows one of the most referenced versions of the Montgomery kP algorithm using Lopez-Dahab projective coordinates (see Algorithm 15 in [2]). The algorithm describes a bitwise processing of the scalar k . The kP algorithm is a sequence of mathematical operations: multiplications, squarings and additions in a finite field. Intermediate and end results have to be written into registers.

Algorithm 1: Montgomery kP using projective Lopez-Dahab coordinates

Input: $k = (k_{l-1} \dots k_1 k_0)_2$ with $k_{l-1} = 1$,
 $P = (x, y)$ is a point of EC over $GF(2^l)$
Output: $kP = (x_1, y_1)$
1: $X_1 \leftarrow x, Z_1 \leftarrow 1, X_2 \leftarrow x^4 + b, Z_2 \leftarrow x^2$
2: **for** $i = l-2$ **downto** 0 **do**
3: **if** $k_i = 1$
4: $T \leftarrow Z_i, Z_i \leftarrow (X_1 Z_2 + X_2 T)^2, X_1 \leftarrow x Z_1 + X_1 X_2 T Z_i$
5: $T \leftarrow X_2, X_2 \leftarrow T^4 + b Z_2^4, Z_2 \leftarrow T^2 Z_2^2$
6: **else**
7: $T \leftarrow Z_2, Z_2 \leftarrow (X_2 Z_1 + X_1 T)^2, X_2 \leftarrow x Z_2 + X_1 X_2 T Z_1$
8: $T \leftarrow X_1, X_1 \leftarrow T^4 + b Z_1^4, Z_1 \leftarrow T^2 Z_1^2$
9: **end if**
10: **end for**
11: $x_1 \leftarrow X_1 / Z_1$
12: $y_1 \leftarrow y + (x + x_1)[(X_1 + x Z_1)(X_2 + x Z_2) + (x^2 + y)(Z_1 Z_2)] / (x Z_1 Z_2)$
13: **return** (x_1, y_1)

Algorithm 1 is a *regular* algorithm: each bit of k , except its most significant bit, is processed with the same type, amount and sequence of operations, independently of the key bit's value. Due to this fact, the Montgomery kP algorithm is referred as resistant against some SCA attacks, such as simple power analysis and simple electromagnetic analysis, see for example [8]. Please note that Algorithm 1 contains many key dependent operations. They are the write-to-register operations in lines 4-5 and 6-7. The assertion about the resistance of the Montgomery kP algorithm against simple analysis attacks is based on the assumption that an attacker cannot distinguish which of the registers is used. The key-dependent register operations are successfully analysed by vertical and differential horizontal address bit attacks [9], [10].

If the regularity principle was not implemented consciously, the implementation can be called “weak implementation”, i.e. its functionality is correct implemented but the power profile for the processing a key bit value ‘0’ can be distinguished from the processing a key bit value ‘1’. In this case the knowledge about the implemented algorithm is sufficient to reveal the processed key successfully. Fig. 1 demonstrates this.

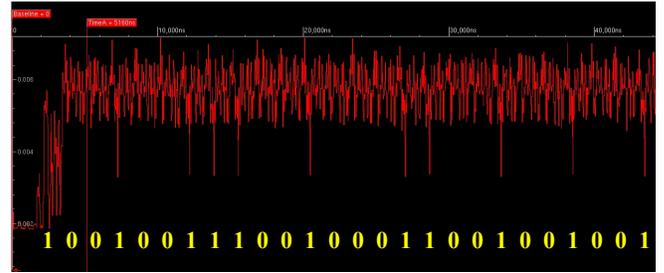


Fig. 1. A part of the simulated power trace of a kP execution of a weak Montgomery kP implementation: the displayed part corresponds to the processing of the first 26 bits of the scalar k .

A power trace for such a weak design was simulated for the IHP 130 nm technology. The clock cycle is 30 ns that corresponds to a clock frequency of about 33 MHz. Fig. 1 shows a part of the simulated power trace of a kP execution. This part corresponds to the processing of the first 26 bits of the scalar k . The revealed bits of the scalar k are shown also.

Inspecting the simulated trace the processed scalar k can be easily revealed using the following assumption: the low amplitude – the dip in the trace – corresponds to the processing of a bit value ‘1’. The dip is not observable when processing of a key bit value ‘0’. The explanation of this weakness is easy: the regularity principal was not guaranteed by this Montgomery kP implementation.

Here we demonstrated that the synchronous IHP kP design selected for the GALS-ification is vulnerable against simple SCA attacks, i.e. the successful revealing of the whole scalar k by simple inspection of the simulated PT is possible. If the assumption that the dip corresponds to a ‘1’ is wrong, the attacker obtains the key via bitwise inversion of the key candidate.

The power trace shown in Fig. 1 was simulated with the time resolution equal to the clock cycle duration, i.e. each clock cycle was represented with only one value that represents the averaged power. Under these conditions the side channel leakage can be

easily detected and used for revealing the key. Fig. 2 demonstrates this. In Fig. 2 – (a) the processing of only the first 4 bits of the scalar k is shown: the time resolution of the yellow graph is 0.01 ns, i.e. resulting in 300 power values per clock cycle; the red graph in the middle was simulated with a time resolution of 0.1 ns, i.e. resulting in 30 values per clock cycle and the red graph at the bottom corresponds to the simulation with the time resolution of 30 ns, i.e. only one – the average – value per clock cycle. Fig. 2 – (b) shows 3 clock cycles from Fig. 2 – (a), zoomed in.

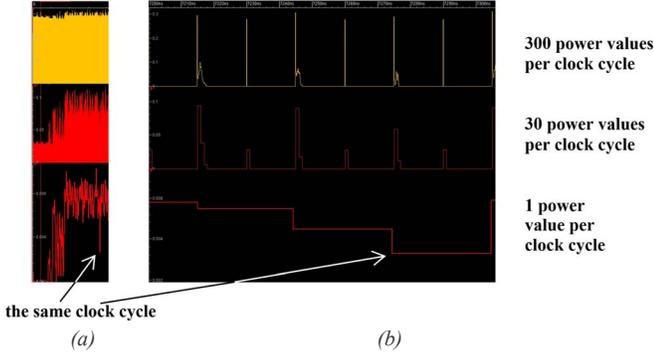


Fig. 2. Power traces of the same kP execution as in Fig. 1 simulated with different time resolutions: 0.1 ns for the yellow graph; 1 ns for the red graph in the middle; 30 ns for the red graph at the bottom.

Please note that to reveal the key only the knowledge about the implemented algorithm and a reasonable representation of the power trace, that was done using different time resolutions for the trace simulation, was sufficient. The same trace representation can be done using the trace compression

III. GALS-IFICATION OF THE SYNCHRONOUS DESIGN

A GALS circuit consists of a set of modules which are locally synchronous, i.e. each module has its own clock. The frequency of the clock signals should differ for the modules. The modules communicate via asynchronous wrappers. A set of modules as a result of a GALS implementation allows to spread the high peak value of the power at the beginning of each clock cycle within the clock cycle or even within a few clock cycles. For example in case of the IHP design described here the field multiplier requires 54 clock cycles for the calculation of 6 field products within a loop of the Montgomery kP algorithm. The field squaring has to be performed only 5 times and needs 2 clock cycles. Thus, the clock frequency for the squaring operation can be – theoretically – about 5 times slower than the one for the multiplication. The same is also true for other operations. In our synchronous circuits the input of each block is connected to the BUS. Only one of all blocks stores or/and processes the value from the BUS. But all blocks consume an internal power if the value on the BUS is changed. This value changes in almost every clock cycle. This causes a high power peak at the beginning of each clock cycle and increases the energy consumption of the kP execution. TABLE I. gives an overview of the power consumption of our synchronous kP design. The values are taken from the reports of the Synopsys synthesis tool.

TABLE I. POWER CONSUMPTION OF OUR SYNCHRONOUS kP DESIGN: AN OVERVIEW.

	switching power	internal power	leak power	total power
kP design	1.23 mW	3.67 mW	2.81 μ W	4.91 mW

It is clearly to be seen that the internal power is about 3 times higher than the switching power. Thus, the GALS-ification of a synchronous circuit seems to be beneficial: it helps to avoid the current spikes at the beginning of each clock cycle and reduces the energy consumption at the same time. In [6] authors propose to apply a GALS-ification as an effective countermeasure against simple power analysis attacks.

The design investigated in [6] is a GALS-ification of the weak synchronous IHP implementation of the Montgomery kP algorithm described in section II. In this weak implementation the processing of a key bit ‘1’ differs from the processing of a key bit ‘0’, whereby the significantly decreased amplitude in synthesized PTs of the synchronous design is the “distinguishable feature” used for the successful revealing of the key. For the GALS-ification the synchronous design was partitioned in 3 blocks:

- field multiplier;
- registers;
- ALU .

Each of these blocks was supplied with its own clock. The frequency was randomized, i.e. not constant in time, but the operation flow was not changed. The details about the implementation of the pausable clocking scheme, with random hopping of clock frequencies can be found in [6]. Here we show schematically the partitioning of our kP design into the blocks for the design’s GALS-ification, the communication between the blocks and the clock signals of the blocks, see Fig. 3.

Fig. 3-(c) shows a part of the simulated traces that corresponds to the processing of the synchronous design for 5 clock cycles. Please note that each local clock signal has its own frequency. The frequencies differ significantly from each other, are independent from each other, and are not constant in time. This increases the execution time of a kP calculation as well as its energy consumption. TABLE II. shows the kP execution time for the synchronous and GALS-ified designs, for comparison. The GALS-ified design needs about double the time for a kP execution in comparison to the synchronous design.

TABLE II. TIME OF THE kP EXECUTION OBTAINED FROM SIMULATED TRACES OF THE SYNCHRONOUS AND THE GALS-IFIED IHP kP DESIGNS.

synchronous design	GALS-ified design	
	<i>Plesiochronous without random hopping clock frequency</i>	<i>Plesiochronous with random hopping clock frequency</i>
390 μ s	580 μ s	890 μ s

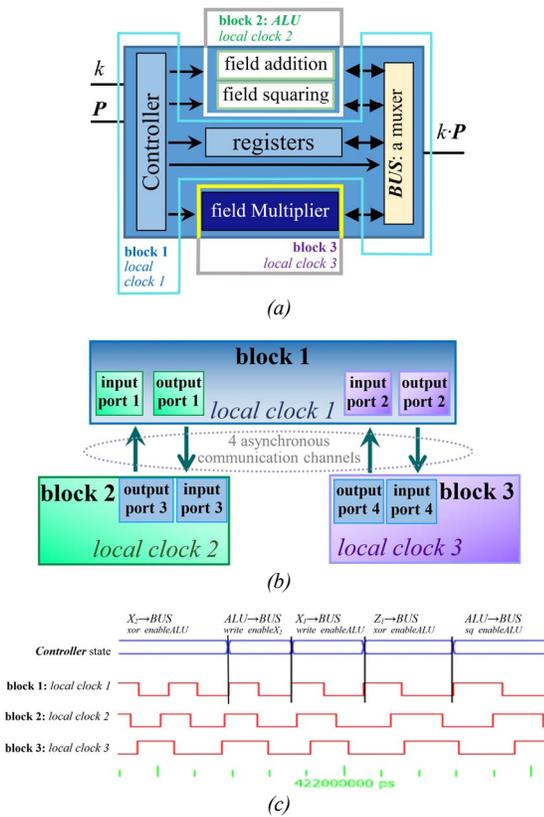


Fig. 3. GALS-ification of the weak synchronous IHP kP design: (a) – partitioning into the blocks for the GALS-ification; (b) – communication between the blocks: each block has its own local clock; four asynchronous communication channels were implemented; (c) – clock signals for the 3 blocks: part of the simulated trace.

Fig. 4 illustrates the relation between the execution time and power consumption during the kP calculation on the example of the simulated power traces of both designs. The simulations were done with the simulation step $\Delta t=1$ ns, i.e. each clock cycle of the synchronous design consists of 30 power values.

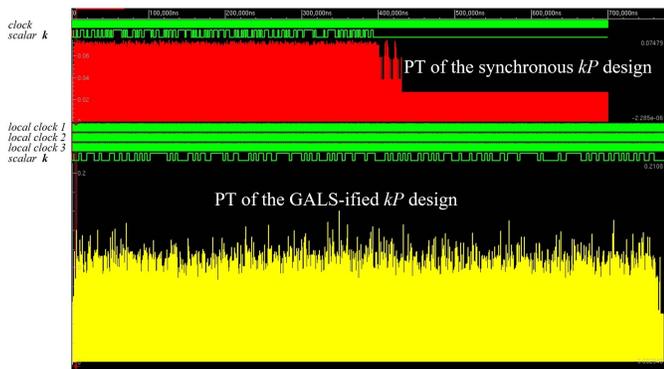


Fig. 4. Simulated traces for the kP execution for the synchronous design (red trace) and for the GALS-ified design (yellow trace). The yellow trace corresponds to the processing of the scalar k only, not to the whole kP execution. Green traces correspond to the clock signals and to the processed scalar k . The time in the diagram is $0 \mu s - 780 \mu s$; the power range for the red trace is $0 W - 0.074 W$ and for the yellow trace it is $0 W - 0.21 W$.

Fig. 5 shows a part of the simulated power traces of the synchronous and the GALS-ified designs, zoomed in. A part at

the beginning of the kP execution is shown. This part corresponds to performing of 3 clock cycles of the synchronous design. For the synchronous design the following traces are shown from top to down: the clock signal, the processed bit value of the scalar k and the power trace of the whole kP design. For the GALS-ified design the following traces are shown: the processed value of the scalar k (it is '0' in the shown part of the trace), the power traces of three local clock cycle generators and the power trace of the whole kP design. The clock signal generator of a block consumes energy to switch its local clock signal. At the end of the generator activity the local clock signal is switched and activates the block, i.e. the block starts to consume the energy. The simulations were done with the simulation step $\Delta t=1$ ns, i.e. each clock cycle of the synchronous design consists of 30 power values.

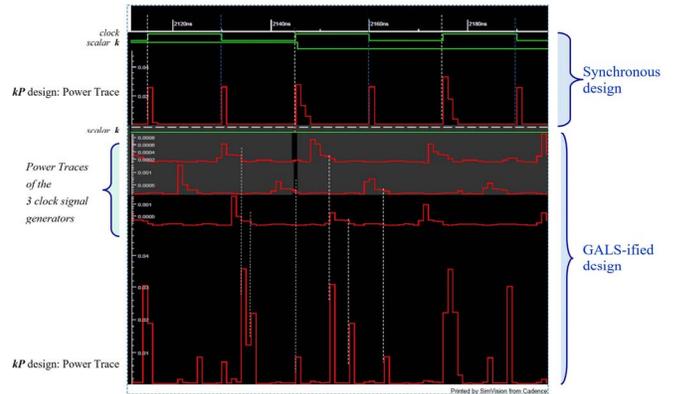


Fig. 5. Simulated power traces of the synchronous and the GALS-ified designs, zoomed in. A part at the beginning of the kP execution is shown. This part corresponds to executing 3 clock cycles of the synchronous design. For the GALS-ified design four power traces are shown: the power traces of the three local clock cycle generators and the power trace of the whole kP design.

Fig. 5 shows clearly that the energy consumption of the GALS-ified kP design is significantly higher than the one of the synchronous design within the same time interval. At the first look the power shape of the GALS-ified design seems to be more complicated than the one of the synchronous design. This can lead to the assumption that the GALS-ification makes the design more resistant against simple SCA attacks, but it is only the first impression that is not true in reality.

The part of the PT of the kP execution shown in Fig. 6 demonstrates why a simulated PT can seem to be complex and how important it is, to set up the simulation parameters correctly.



Fig. 6. Power trace of the kP execution of the GALS-ified design simulated with the three different simulation steps: $\Delta t=0.1$ ns for the top PT, $\Delta t=1$ ns for the middle PT, $\Delta t=30$ ns for the bottom PT. All traces correspond to the same part of the kP execution.

The three power traces in Fig. 6 correspond to the same part of the same kP execution. Only the simulation step differs for all 3 traces. The top trace was simulated with the simulation step $\Delta t=0.1$ ns, the trace in the middle of the diagram was simulated with the simulation step $\Delta t=1$ ns that is the same simulation step as in Fig. 5 and the bottom trace was simulated with the simulation step $\Delta t=30$ ns.

For designers it is important to understand that the resolution with that the simulation in run has a significant impact on the result of the analysis, i.e. that for the design discussed here the simulation step $\Delta t=30$ ns may not be applied if the goal is to evaluate the resistance of the design against SCA attacks. The design contains three clock generators. Each clock generator generates a local clock signal with the period between 20 ns and 40 ns. To see the activity of functional blocks with these clock signals the simulation step has to be significantly smaller than 20 ns. This is the difference that we can see in Fig. 6 if we compare the bottom PT simulated with $\Delta t=30$ ns with the middle PT simulated with $\Delta t=1$ ns. The best “resolution” has the PT simulated with $\Delta t=0.1$ ns, see the top PT in the diagram. In this diagram the activity of the block with the highest energy consumption can be clearly seen. The fact that the most consuming block in a kP design is the field multiplier is well-known. Thus, if an attacker can see the trace that corresponds to the activity of only the field multiplier within the whole kP operation, he can concentrate on attacks specialized on exploiting different multiplier weaknesses. The same is also true for other blocks, for example the block with the registers. The GALS-ification described here can help an attacker to perform a successful analysis of the activity of registers. Fig. 7 demonstrates how the common knowledge about the power consumption of the mathematical operations and registers allows to understand, what we “see” in the PT. The letter **M** marks the peak that corresponds to the activity of the field multiplier, that is the biggest and most energy consuming block of the design, i.e. it is *block 3* of the GALS-ified kP design. It is more complicated to understand, which peaks correspond to the activity of the block with the registers (*block 1* in the GALS-ified kP design) and of the *ALU* (*block 2*). The common knowledge that can help to select the activity of the registers is the fact that their activity is very short in time. In Fig. 7 letter **R** denotes the peak that corresponds to the activity of the block with registers and letter **A** corresponds to the activity of the *ALU*. We did this separation using the knowledge of designers about the activity of the clock signal generators, i.e. comparing PTs of clock generators with PTs of the design blocks. But the short pulse duration allows to distinguish the power profile of the *ALU* from the one of the registers, even for attackers.

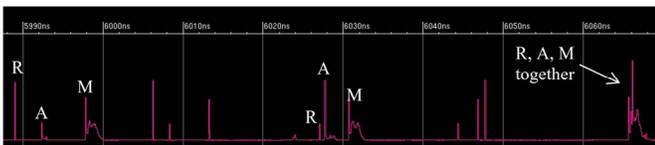


Fig. 7. A part of the power trace of the kP execution of the GALS-ified design simulated with $\Delta t=0.1$ ns. The power profiles of the three design blocks differ significantly from each other and can be distinguished using common knowledge about the implementation details of mathematical operations.

Please note that for the GALS-ification described here the following is true: each pulse in the power trace of the synchronous design corresponds to a “triplett” in the power trace of the GALS-ified design. Moreover, the pulse that corresponds to the multiplier can be clearly identified in each triplet. Due to this fact the distinguisher that we used for the successful SPA on the synchronous design has to be successful also when attacking the GALS-ified design. This means that each clock cycle with low power in the synchronous design (that is the distinguishing marker for the processing of a key bit value ‘1’) has to correspond to a triplett where usually the biggest power consumption is now significantly lower than in other tripletts. Thus the key can be easily revealed attacking the GALS-ified design too. Fig. 8 demonstrates this.

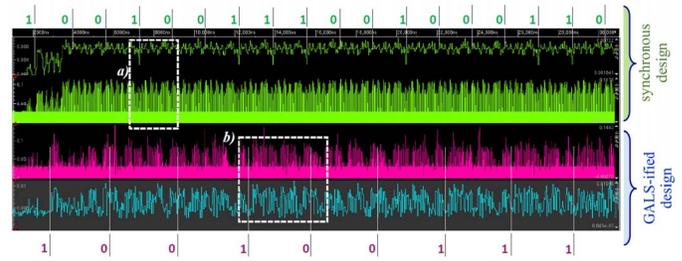


Fig. 8. A part of power traces of the kP execution of the synchronous and the GALS-ified designs. The top (green) and the bottom (blue) traces in the diagram were simulated with $\Delta t=30$ ns. Both traces in the middle (green and magenta) were simulated with $\Delta t=1$ ns. The processed key bit values for the synchronous design are shown at the top of the diagram, see green marked sequence of ‘1’ and ‘0’. The processed key bit values for the GALS-ified design are shown at the bottom of the diagram, see magenta marked sequence of ‘1’ and ‘0’. The part of the traces marked with white dashed rectangles *a)* and *b)* corresponds to the processing of the same key bits of the scalar k .

The two green traces in Fig. 8 are power traces of the kP execution of the synchronous design. The top green trace was simulated with the simulation step $\Delta t=30$ ns that is equal to the period of the clock signal. In this case each value in the simulated trace represents the energy consumption within the corresponding clock cycle. Due to the fact that the distinguishing marker is the low energy consumption during a clock cycle, the applied simulation step helps to “see” these clock cycles. Thus, each dip in the trace corresponds to the processing of a key bit ‘1’. The processed key bit values for the synchronous design are shown at the top of the diagram, see green marked sequence of ‘1’ and ‘0’. The other green trace was simulated with the simulation step $\Delta t=1$ ns. The distinguishing marker for the processing of a key bit value ‘1’ is now not easy to “see”, but this doesn’t mean that the SPA doesn’t work.

The magenta trace in Fig. 8 is the power trace of the kP execution of the GALS-ified design simulated with the simulation step $\Delta t=1$ ns. In this curve groups of pulses with high amplitude can be seen. These are short groups containing 10 “high” pulses and long groups containing 20 “high” pulses. The high pulses correspond to the field product calculation steps. The groups with 10 high pulses correspond to one field multiplication. The groups with 20 high pulses are two field multiplications calculated immediately one after the other. The violet line in Fig. 9 shows the part of the magenta trace identified by the rectangle *b)* in Fig. 8, zoomed in. The red line is the PT of the partial multiplier that is a part of the field multiplier. Due

to the fact that the partial multiplier consists only of combinatorial gates, it reacts on the change of the partial multiplicands only. It is clearly to be seen that the violet trace has much more pulses than the red trace but the “high” pulses in the violet line correspond to the field product calculations.

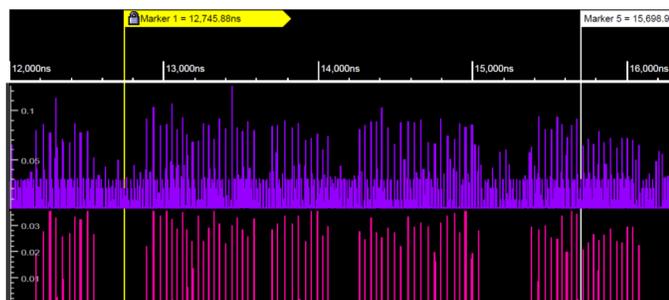


Fig. 9. A part of power traces of the kP execution of the whole GALS-ified design (violet trace) and its Partial Multiplier (red trace). Two time markers show the end of the time slots for the processing a key bit. The yellow time marker corresponds to the end of the processing of a key bit value ‘1’. The white time marker corresponds to the end of the processing of a key bit value ‘0’.

Two time markers in Fig. 9 show the end of the time slots of processing a key bit. The yellow time marker corresponds to the end of the processing of a key bit value ‘1’. The white time marker corresponds to the end of the processing of a key bit value ‘0’. The fact that the design consumes much less energy at the end of the processing a key bit ‘1’ can successfully be exploited for revealing the key also for the GALS-ified design. At the end of the processing of a ‘1’ (see yellow time marker) the waiting of the multiplier between two field multiplication is long, i.e. we see a gap between groups with high peaks. At the end of the processing of a ‘0’ the field multiplier doesn’t wait, i.e. we see the white time marker in the middle of a group with high peaks. Using this “new” distinguisher the key can be revealed successfully even for the trace simulated with the rough simulation step of 30 ns, see Fig. 9. This shows clearly that a straight forward GALS-ification of a synchronous ECC design that is vulnerable to SPA is also vulnerable to SPA.

IV. CONCLUSIONS

GALS-ification can be a good countermeasure, but the GALS-ification may not be straight forward, i.e. it needs to be more sophisticated than just using existing blocks. Otherwise the weaknesses of a synchronous design will survive the transformation leading to the fact that the GALS-ified design is also vulnerable. This is due to the fact that in a GALS-ified design the activity of the single blocks can be observable. This makes the GALS-ified designs even more vulnerable to a wide spectrum of SCA attacks. When a GALS-ified design is analysed in order to verify its vulnerability a reasonable

representation of the simulation results and even more important a proper definition of the simulations steps is of utmost importance. Otherwise the results may lead to misinterpretations especially to ones emphasising resistance against SCA.

In order to really improve the SCA resistance of a design using GALS-ification a sophisticated strategy is paramount. A potential approach is splitting the design in a lot of blocks with similar power shapes, so that the power profiles of different operations are no longer distinguishable from each other.

ACKNOWLEDGMENT

The authors would like to thank the Dr. Milos Krstic from System department of IHP for providing the GALS-ified version of IHP ECC design.

REFERENCES

- [1] Federal Information Processing Standard (FIPS) 186-4, Digital Signature Standard; Request for Comments on the NIST-Recommended Elliptic Curves: 2015.
- [2] López, J., Dahab, R.: Fast multiplication on elliptic curves over $GF(2^m)$ without precomputation. In: Koç, Ç.K. and Paar, C. (eds.) Cryptographic Hardware and Embedded Systems. pp. 316–327. Springer Berlin Heidelberg, Berlin, Heidelberg (1999).
- [3] Darrel Hankerson, Alfred Menezes, Scott Vanstone: Guide to Elliptic Curve Cryptography, Springer-Verlag New York, Inc., 2004, ISBN 0-387-95273-X
- [4] I. Kabin, D. Kreiser, Z. Dyka, and P. Langendoerfer, “FPGA Implementation of ECC: Low-Cost Countermeasure against Horizontal Bus and Address-Bit SCA,” in 2018 International Conference on ReConfigurable Computing and FPGAs (ReConFig), 2018, pp. 1–7.
- [5] A. Bauer, E. Jaulmes, E. Prouff, and J. Wild, “Horizontal Collision Correlation Attack on Elliptic Curves”, in *SAC* 2013, pp. 553–570.
- [6] Xin Fan, S. Peter and M. Krstic: GALS design of ECC against side-channel attacks — A comparative study. Proceedings of 24th International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS), Sept. 29-Oct. 1 2014, pp. 1 – 6, IEEE
- [7] P. L. Montgomery, “Speeding the Pollard and elliptic curve methods of factorization”, *Mathematics of Computation*, 48s, 177 (Jan. 1987), p. 243–243.
- [8] Junfeng Fan, Xu Guo, Elke De Mulder, Patrick Schaumont, Bart Preneel, Ingrid Verbauwhede: State-of-the-art of Secure ECC Implementations: A Survey on Known Side-channel Attacks and Countermeasures, Proceedings of the 2010 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST 2010), 13-14 June 2010, Anaheim Convention Center, California, USA. IEEE Computer Society, pp.76-87, Jul 2010
- [9] Itoh, K., Izu, T., Takenaka, M.: Address-Bit Differential Power Analysis of Cryptographic Schemes OK-ECDH and OK-ECDSA. In: Cryptographic Hardware and Embedded Systems - CHES 2002. pp. 129–143. Springer, Berlin, Heidelberg (2002).
- [10] I. Kabin, Z. Dyka, D. Kreiser, and P. Langendoerfer, “Horizontal Address-Bit DEMA against ECDSA,” in 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 2018, pp. 1–7.

Unit Regression Test Selection Mechanism Based on Hashing Algorithm

Melikyan Vazgen Sh., Hakobyan Hovhannes H., Kaplanyan Taron K., Momjyan Arsen M.

Synopsys Armenia Educational Department

Synopsys Armenia CJSC

Yerevan, Armenia

vazgenm@synopsys.com, hhovo@synopsys.com, tkaplan@synopsys.com, momjyan@synopsys.com

Abstract—The regression test selection mechanism is represented, which leads to a decrease in the overall testing time. Provided the detailed description of the software, which selects and classifies the tests, as well as the initial results of the work done with the software, based on the safe regression test select mechanism.

Keywords—RTS, software, unit testing, hashing

I. INTRODUCTION

Software systems are regularly modified during their formation and operation periods for several reasons: correction of errors, new functionality addition or performance speed boost. After the software modification, the updated version must be tested to ensure that the applied changes do not have a negative impact on the current version of the software. The aim of a regression test is to check the quality of the new version of the software after making some changes in it. In most cases it is not possible to perform a complete regression test due to the shortage of time, as in practice a new version's releasing cycle is quite short. In order to reduce the regression test time, researchers develop methods that will reduce the "cost" of regression testing [1-2]. One of the methodologies is to have a T_i tests' set, which is generated to test P_i version of the software, and to use the same tests' set also for next P_{i+1} version. However, it is recommended to use the "selective regression test" approach. Using RTS (Regression Test Selection) method will increase the re-testing efficiency. This method suggests separating a T_i' subset from T_i set and use it for P_{i+1} version testing [1-3]. This method is considered to be safe, as the $(T_i \setminus T_i')$ test results will be the same for both P_i and P_{i+1} versions, and using T_i' set (number of tests are less than in T_i set) will formulate the same result with the usage of less resources [3][4]. The graphical representation of RTS process presented in Fig. 1.

Several RTS methods use the source code of the software, from which they collect the coverage information. Coverage information, such as requirements, branches is collected while testing P_i version with T_i set of tests and is used for generating the T_i' tests' set for P_{i+1} version.

In this case, testing only the functionality that is modified by the implemented changes, instead of executing a full-functionality test, will shorten the deadlines of the new version release. In order to reach this, usually random tests selection method is applied, meaning that tests are designed randomly or

small amount of regression tests is done. In both cases, a full-functionality regression test is not performed, which can lead to some unnoticed errors (bugs) in the software, resulting to an inaccurate performance of the whole system. Thus, before making the release, regression tests needs to be conducted the way, that all the modified parts are properly tested. In practice, there is a developed method of regression test selection, which is designed to identify the modified sections and perform tests related to those parts. The aim of this work is to create an RTS mechanism, which will be used in practice. According to this method, instead of performing a complete regression test, it is possible to select a subset of tests. The subset is selected so that the results of the complete regression test and the subset are the same. This leads to the decrease of testing time. The advantage of this mechanism is that it is adaptable to existing technologies. It selects a subset of tests, which are related only to the changed parts.

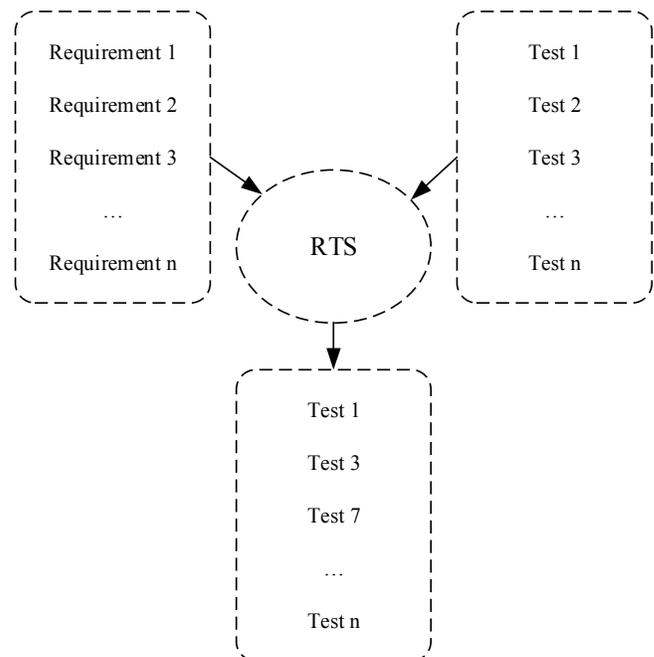


Fig. 1. RTS process

II. SELECTIVE REGRESSION TESTING

This section provides the selective regression testing description and an example that will make easier to understand the work of the RTS.

Figure 2 shows a small part of the program represented in the Python scripting language. The provided example includes a schemes' database (DB) and several functionalities (requirements). The 4 requirements are r1 (DB_connection), r2 (validate_user), r3 (Get_permissions), r4 (check_cell_view): Software consists of two parts: "User" and "DB". "User" displays the user's name, email address and more. "DB" displays the schemes' database, schemes' descriptions, availability, etc.

DB_connection function performs the r1 requirement. This function checks the database access connection. If the connection was successful and user was able to sign in, than he should be able to perform actions that are allowed to that particular user. Otherwise, user gets and error message "Error: cannot connect to DB".

Validate_user function performs the r2 requirement, which checks the user's info by sending a request to database to get information about all the users. If the requested user's name is found in DB, then the email check is performed. This check is done to differentiate users with the same name.

Get_permissions function performs the r3 requirement. This function as an argument uses the results of the validate_user function. If the user does not have the permissions and the check brings negative results, then the user get the following message: "Error: Invalid user specified". In the opposite case, the requestor gets the list of permissions for the specified user.

Check_cell_view function performs the r3 requirement. This function as an argument uses the scheme description type. If the scheme was not found in the database, then the user gets the following message:

"Error: Cell cell.name does not exists in DB". If the scheme is detected in DB, the list of scheme descriptions is generated, where a search is done to find initially selected description.

Each requirement (functionality) is described by a critical degree, which shows the importance of the requirements.

Table 1 represents the details of those r1-r4 requirements. Here the critical degrees are tentatively set as numbers, the higher the number the higher the importance of the requirement check.

TABLE I. SYSTEM REQUIREMENTS AND CRITICAL DEGREES

Requirement	Critical degree	Description
r1	3	Check: DB connection functionality
r2	4	Check: User's information checking unit
r3	2	Check: User permissions checking unit
r4	1	Check: Schemes' descriptions checking unit

```
class DB:
    def __init__(self):
        self.user = User()
        self.error_num = 0
        self.cell_list = list()
    ...

class User:
    def __init__(self):
        self.name = ""
        self.mail = ""
    ...
```

Fig. 2. Class definitions

```
def DB_connection(self):
    try:
        db = MySQLdb.connect(self.server, self.user,
                               self.passwd,
                               self.schema)
        cursor = db.cursor()
        cursor.execute("SELECT VERSION()")
        results = cursor.fetchone()
        if results:
            return True
        else:
            return False
    except MySQLdb.Error:
        print "Error: Cannot connect to DB."
        self.error_num = 1
    return False
```

Fig. 3. System Requirement 1

```
def validate_user(self, user):
    try:
        user_list = self.db.get_users()
        if len(user_list) > 1000: #change 1.1 if
            len(user_list) > 2000:
                print "Warning: Users in Cell DB
                should be less than 1000" #change 1.2 .. less than 2000
            except MySQLdb.Error:
                raise ValueError("Error: Cannot gat user names
                from DB.")
            error_num = 2
            if user.name in user_list:
                if re.match('^[_a-z0-9-]+(\\.[_a-z0-9-]+)*@[a-z0-9-]+(\\.[a-z0-9-]+)*\\.([a-z]{2,4})$', user.email):
                    return True
                else:
                    raise ValueError("Error: User
                    user.name mail user.email has incorrect form.")
                    error_num = 3
            else:
                raise ValueError("Error: User user.name does not
                in DB user list.")
                error_num = 4
```

Fig. 4. System Requirement 2

Two changes are made in the software. First change is done in validate user function, where the number of users in database is changed from 1000 to 2000. Second change is done in get_permissions function. In the previous version the result was a printed message. After the upgrade “raise” command is used, which not only prints the message but also stops the work of the program.

```
def get_permissions(self, user):
    try:
        validate_user(user)
    except ValueError:
        print "Error: Invalid user specified"
    try:
        perms = self.db.get_permissions(user)
    except:
        print "Error: Cannot get permissions for user.name
user."
        //change 2
        raise ValueError('Error: Cannot get permissions for
user.name user.')
```

Fig. 5. System Requirement 3

```
def check_cell_view(self, cell, view):
    try:
        found = self.db.cell_exists(cell)
    except:
        print "Error: Cell cell.name does not exists in DB."
    try:
        views = self.db.get_cell_views(found)
    except:
        print "Error: Cannot gat view list of cell.name cell."
    if view in views:
        return True
    else:
        print "View view does not exists for cell.name cell."
        return False
```

Fig. 6. System Requirement 4

TABLE II. PERFORMED TESTS

Test Cases	Names
t1	test_DB_connection
t2	test_validate_user
t3	test_get_permissions
t4	test_check_cell_view

Table 2 shows T set of tests, which includes tests from t1 to t2. Those test cases need to be performed when a change is implemented. To make this example clear, each requirement has one corresponding test case, but in fact, test cases can be more than one. In order to create compliance between test cases and functions the idea of requirement traceability matrix [5] is used. It is a document, which describes the system requirements and the corresponding test cases. The main aim of this document is to make sure that each component of the software has at least one test case, so the overall system is tested.

In Table 3, the 0 number shows that the test case does not affect the requirement. If the number is more than 0, then the test affects the requirement. So for r1 required test case is t1, with critical degree equal to 3: This number is used for further test collection and for prior bug detection.

TABLE III. SYSTEM REQUIREMENTS AND TEST CASES WITH THEIR CRITICAL VALUES

Requirements	Test cases			
	t1	t2	t3	t4
r1	3	0	0	0
r2	0	4	0	0
r3	0	0	2	0
r4	0	0	0	1

III. SAFE REGRESSION TEST SELECTION

RTS methods are divided into two groups: Safe regression tests and unsafe regression tests.

- In case of safe regression testing T_i and T'_i test results are the same for the software P_i and P_{i+1} versions. So, by running T'_i set of tests all the changes are being tested, which saves time.
- In case of unsafe regression testing T_i and T'_i test results are different for the software P_i and P_{i+1} versions, which means that T'_i set of tests is not enough for full testing of P_{i+1} version [4].

To define the safe regression testing method the DEJAVOO software has been analyzed. This software creates control-flow graphs for P_{orig} and P_{mod} versions of the software. Then it skips the 2 mentioned stages, which can have the following three values: empty, right, false. The mechanism runs the “search by depth” algorithm to find “dangerous” edges.

Edges that differs in the P_{mod} version are considered dangerous. The test set which was ran on P_{orig} version needs to be run also on P_{mod} version, as the results could change. To understand how DEJAVOO works, let’s look at validate_user function’s current(v0) and modified(v1) versions. The control-flow graphs are presented below in Fig. 3.

Software skips the graphs from the first node. The commands performed in the function are numbered to facilitate further actions. Until reaching the third command the skip result will not change, as in both v0 and v1 versions there is no command change. Reaching the 3rd command, software detects a difference compared to the original version. The RGS algorithm marks that part of the code as “dangerous”, after which the software uses the initially provided coverage matrix, which describes the connection between the tests and the commands of the function.

Table 4 presents the coverage matrix of a function which has m edges(commands) and n test cases. When finding all the “dangerous” edges, software performs a search in coverage matrix. All the found tests are added to T'_i test set, in order to test the P_{i+1} version.

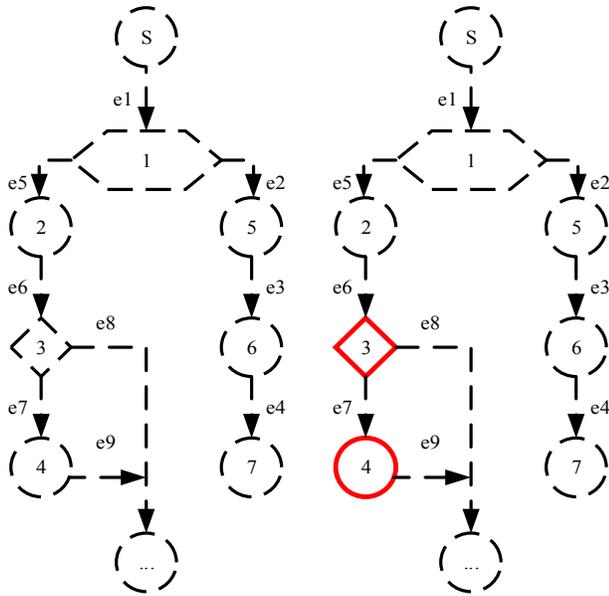


Fig. 7. The control flow graph

TABLE IV. COVERAGE MATRIX

Edge	Test Cases					
	t_1	t_2	t_3	...	t_{n-1}	t_n
e_1	1	0	1	...	1	0
e_2	0	0	0	...	1	1
e_3	1	0	1	...	1	1
...
e_{m-1}	0	0	0	...	0	1
e_m	1	0	0	...	0	1

IV. DEVELOPED MECHANISM

The `validate_user` function is taken as an example to demonstrate how the developed algorithm works. Below the `validate_user`'s initial and modified versions.

```
def validate_user(self, user):
try:
    user_list = self.db.get_users()
    if len(user_list) > 1000:
        print "Warning: Users in Cell DB should be less than 1000"
except MySQLdb.Error:
    raise ValueError("Error: Cannot gat user names from DB.")
    error_num = 2
...
```

Fig. 8. Function example: original requirement

```
def validate_user(self, user):
try:
    user_list = self.db.get_users()
    if len(user_list) > 2000:
        print "Warning: Users in Cell DB should be less
than 2000"
except MySQLdb.Error:
    raise ValueError("Error: Cannot gat user names from DB.")
    error_num = 2
...
```

Fig. 9. Function example: changed requirement

Figure 10 shows the developed RTS mechanism's block scheme and the inputs and outputs of each block.

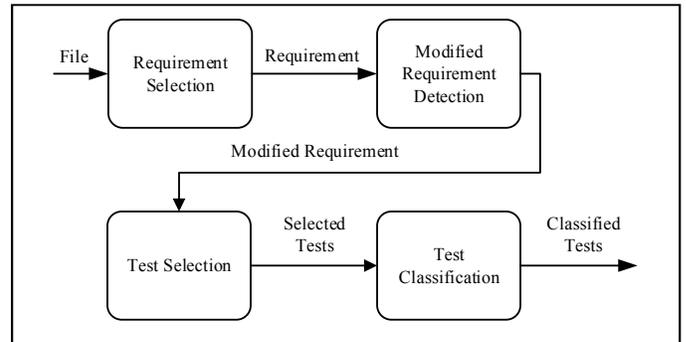


Fig. 10. Block diagram of the developed RTS mechanism

The "Requirement Selection" block uses as an input a file generated for the new version of the software. Getting the file, this block searches and finds all the requirements.

The "Modified Requirement Detection" block uses as an input the list of requirements from the first block. Then each command line from the requirement is highlighted using the MD5 algorithm. Retrieved hash values are compared to current hash values. If those values differ, that there were made some changes in that requirement, so the corresponding tests needs to be performed for the next version of the software.

The "Test Selection" block gets the modified requirements' list and by using the traceability matrix detects the required tests.

The final "Test Classification" block sorts the selected tests by critical degrees.

The test selection consists of 4 main steps:

1. On the first step hash value is calculated for all the command of the requirement, and the values are stored in the `pair_list` list.
2. On the second step the hash values of each pair is compared to the initial one. If the values differ, then that pair is stored in `change_list` list.
3. On the third step the algorithm searches for corresponding tests for the stored changes in the `change_list` and adds them to `must_run` list.
4. On the forth step all the tests in `must_run` list are sorted by their critical degree.

In Fig. 11 provided pseudo-code of the test selection.

```

Step 1:
foreach r ∈ Requirements
  foreach command_line ∈ r
    hash_value = hash_value + hash(command_line)
  endfor
pair_list.append(r,hash_value)

Step 2:
foreach pair ∈ pair_list
  diff = compare(pair[hash_value], current_hash)
  if (diff)
    change_list.append(pair)
  endif
endfor

Step 3:
foreach change ∈ change_list
  change_tests = find_tests(change)
  must_run.append(chang_tests)
endfor

Step 4:
prioritize_tests(must_run)

```

Fig. 11. Test selection pseudo-code.

Thus, a safe RTS mechanism was developed and implemented based on hashing algorithm. It allows to select a subset of tests that are related to changes made in the software system. Performing testing based on the subset of selected tests, instead of performing a complete regression test before the release of the version of the software brings to reduction of the overall testing time.

V. RESULTS

The Table 5 represents the experimental results of developed mechanism. It is showing the execution time depending on different number of functions. Each function contains 500+ lines body. Provided results represented in ms. In Fig. 12 presented the average execution time dependent on number of functions.

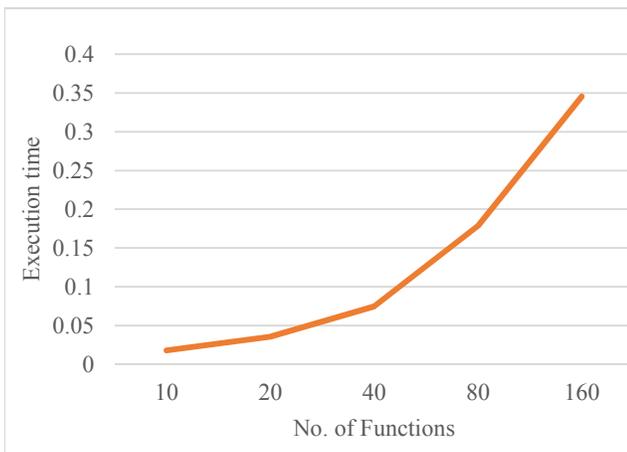


Fig. 12. Execution time dependency on number of functions

TABLE V. EXECUTION TIME

Runs \ Func.	Run 1	Run 2	Run 3	Run 4	Run 5
10	0,0183	0,0198	0,0098	0,0206	0,0216
20	0,0380	0,0265	0,0459	0,0335	0,0360
40	0,0773	0,0718	0,0884	0,0673	0,0685
80	0,1989	0,1320	0,2184	0,1650	0,1774
160	0,3474	0,3783	0,2803	0,3656	0,3554

The analysis performed on Intel core i5 6th generation CPU with 4 cores and 2.4Ghz operating frequency. It is used 64-bit operating system.

VI. CONCLUSION

A new test selection approach is suggested. The approach based on RTS techniques, which uses traceability matrix and system requirements. The suggested RTS method uses hashing algorithm, which preventing storing high amount of data. Execution time analysis results is provided. In this method it is not used complex data structures(for example, storing data in graph and traversing over it), which is common for RTS processes. Due to that approach it is reduced data usage and overall execution time.

REFERENCES

- [1] P. K. Chittimalli and M. J. Harrold. "Re-computing coverage information to assist regression testing" In International Conference on Software Maintenance (ICSM 2007), October 2007, pp. 164-173.
- [2] M. J. Harrold. "Testing evolving software" In The Journal of Systems and Software 47, 1999, pp. 173-181.
- [3] P. K. Chittimalli. "Regression test selection on system requirements" Conference Paper, January 2008, pp. 87-96.
- [4] A. Orso, N. Shi, and M. J. Harrold. "Scaling regression testing to large software systems" In Proceedings of the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Newport Beach, CA, October 2004, 11 P.
- [5] L. Briand, Y. Labiche, and S. He. "Automating regression test selection based on UML designs" Information and Software Technology, Vol 51, No. 1 pp. 16-30, January 2009.
- [6] P. Chittimalli and M. Harrold. "Regression test selection on system requirements" In ISEC '08. Proceedings of the 1st conference on India software engineering conference, pp. 87-96, 2008.
- [7] <https://www.guru99.com/traceability-matrix.html?fbclid=IwAR0FtRFVVowTWQosjN9f3DqEVIEYW3Z9ssD2GTvGJQLc6hSdR4-JRb>

Qubit Fault Detection in SoC Logic

Mikhail Karavay
*V.A. Trapeznikov Institute of
 Control Sciences of Russian
 Academy of Sciences*
 Moscow, Russia
 mkaravay@yandex.ru

Vladimir Hahanov
*Design Automation Department
 Kharkov National University of
 Radioelectronics*
 Kharkov, Ukraine
 hahanov@icloud.com

Eugenia Litvinova
*Design Automation Department
 Kharkov National University of
 Radioelectronics*
 Kharkov, Ukraine
 litvinova_eugenia@icloud.com

Hanna Khakhanova
*Design Automation Department
 Kharkov National University of
 Radioelectronics*
 Kharkov, Ukraine
 ann.hahanova@gmail.com

Irina Hahanova
*Design Automation Department
 Kharkov National University of Radioelectronics*
 Kharkov, Ukraine
 irina.hahanova@nure.ua

Abstract— A memory-driven technology for diagnosing and testing digital systems-on-chips is proposed, which allows increasing the performance of design and test processes by an order of magnitude. The technology differs from the classical theory in the absence of traditional logic for the synthesis and analysis of computing devices and the use of qubit data structures of quantum computing. All the functionalities of computer architectures are designed on the basis of memory components, which implement sequential primitives and combinational circuits. The model of a universal digital primitive excludes a truth table and is represented by a qubit binary vector that adequately describes the behavior of any functionality of combinational or sequential type. Methods for the synthesis of functionalities based on qubit vectors in memory are considered, which compactly describe the behavior of computer components. Methods for analyzing digital architectures based on qubit vectors are proposed, which make it possible to effectively simulate the behavior of computing devices by using only read-write transactions in memory. Methods for testing digital systems defined by qubit vector structures are considered, which make it possible to effectively generate tests and simulate faults for computing devices by using parallel vector logic operations. The effectiveness of the proposed models and methods for the synthesis and analysis of digital systems is metrically evaluated by performance parameters, memory hardware costs, and also the quality of the designed processes and products.

Keywords— *design and test, qubit data structures, unitary coding, qubit-driven diagnosis, logic functions, quantum computing*

I. QUBIT DATA STRUCTURES OF QUANTUM PRIMITIVES

The goal is the implementation of quantum computing technologies in the practice of improving the performance of Design and Test models and methods by organizing parallel computational processes for the synthesis and analysis of tests, and also conducting diagnostic experiments based on the use of qubit data structures. Objectives are the following: 1) Consideration of the advantages of the qubit description of logical elements. 2) Creating a macromodel for testing using xor relations between the components, which form the processes of synthesis and analysis of computing devices. 3) Scaling the advantages of the Deutsch algorithm for diagnosing logical functions of a finite number of variables. 4) Development of the

qubit method for diagnosing stack-at faults of digital circuits. 5) Development of memory-driven computing architectures for parallel solving Design and Test problems.

The features of the theoretical foundations of quantum circuitry are associated with the use of linear algebra for unitary coding of the states of logic elements, which provides the possibility of space-time parallelism in solving actual problems of testing digital systems. The main theoretical points necessary for the presentation of models and methods for the synthesis and analysis of discrete devices are considered on several examples [1-7]: 1) Quantum gates transform qubits using a unitary linear reversible operator, and also in the form of a matrix representation of quantum gates [1-2]: $|\psi\rangle, |\chi\rangle \in \mathbb{C}^2$, $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

It is convenient to use the Dirac form: $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, $|\chi\rangle = \gamma|0\rangle + \delta|1\rangle$, $A|\psi\rangle = |\chi\rangle$, which is easily transformed into a matrix form: $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \gamma \\ \delta \end{pmatrix}$. An example of a quantum NOT gate defined by a unitary matrix: $X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. It is easy to see that the last expression is easily converted to the form:

$$X(\alpha|0\rangle + \beta|1\rangle) = \beta|0\rangle + \alpha|1\rangle,$$

where the NOT gate is a switch between states $|0\rangle$ and $|1\rangle$. Another example of quantum gate described by unitary matrices is the following:

$$Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

which act on the qubit in accordance with the expressions:

$$Y(\alpha|0\rangle + \beta|1\rangle) = -i(-\beta|0\rangle + \alpha|1\rangle),$$

$$Z(\alpha|0\rangle + \beta|1\rangle) = \alpha|0\rangle - \beta|1\rangle,$$

where X, Y, Z are the Pauli matrices. Hadamard gate performs a reversible operation of forming a superposition of states and is determined by the unitary matrix [4, 5, 7]:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The following expression is true $X^2 = Y^2 = Z^2 = H^2 = I$, which means that each gate X, Y, Z, H is determined by the square root of the unit matrix and the corresponding quantum gate I. In a multi-qubit gate, there must be the same number of qubits at the input and output. 2) Two-qubit gates correspond to the rotation operations in the Hilbert space of two interacting qubits, which cannot be represented as a direct product of independent single-qubit operations [4, 5, 7]. The main two-qubit gate is a reversible controllable CNOT inverter with two input and two output qubits (Fig. 1), which operates in accordance with the following expressions:

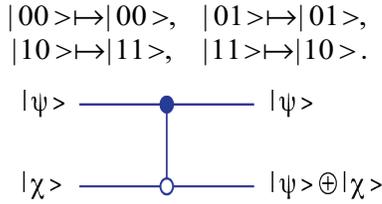


Fig. 1. Controllable inverter

When $|\psi\rangle=|0\rangle$, then $|\psi\rangle\oplus|\chi\rangle = |\chi\rangle$ for $|\psi\rangle=|1\rangle$ is true $|\psi\rangle\oplus|\chi\rangle = X|\chi\rangle$.

Controllable CNOT gate is described by a matrix:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ \delta \\ \gamma \end{pmatrix},$$

where $|\psi\rangle=\alpha|0\rangle + \beta|1\rangle$, $|\chi\rangle = \gamma|0\rangle + \delta|1\rangle$.

Qubit $|\chi\rangle$ is the control one, qubit 1 – controllable one, on which the NOT operation is performed, provided that the first qubit is in the state $|1\rangle$.

3) The two-qubit SWAP qubit state exchange operator can be implemented by successively performing three CNOT operations (Fig. 2) [4, 5, 7] and is described by the matrix:

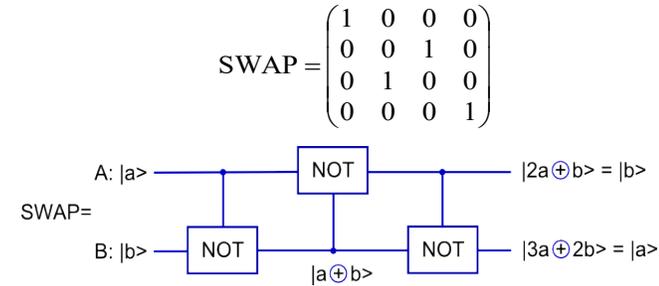


Fig. 2. SWAP operator

4) The three-qubit Toffoli gate (CCNOT) is shown in Fig. 3, contains two control qubits A and B, and also one controllable one C [6].

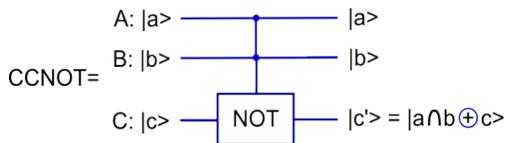


Fig. 3. Toffoli gate

The Toffoli gate is described by 8x8 matrix in basic states $|0,0,0\rangle, |0,0,1\rangle, |0,1,0\rangle, |1,0,0\rangle, |0,1,1\rangle, |1,0,1\rangle, |1,1,0\rangle, |1,1,1\rangle$:

$$\text{CCNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

This operation is implemented in the form of five two-qubit operations. In general, the Toffoli N-bit gate is described by the expression:

$$(k_1, k_2, \dots, k_n) \rightarrow (k_1, k_2, \dots, (k_1, k_2, \dots, k_{n-1}) \oplus k_n)$$

The NOT element is a special case of the Toffoli gate for which $n = 1$, and CNOT is the case when $n = 2$. An extended $n+1$ -bit Toffoli gate (ETG) with two controllable lines is shown in Fig. 4. The ETG circuit has input and output qubit vectors. The first $n-1$ bits are the control ones, the last two are controllable.

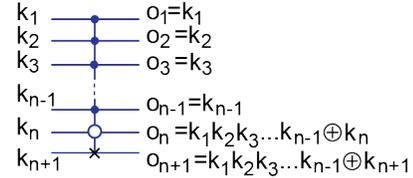


Fig. 4. $n+1$ -bit Toffoli gate

To improve computational performance by solving testing and diagnostics problems in parallel, the following properties of quantum logic and qubit data structures should be used: 1) Unitary data coding, which makes it possible to perform logical operations in parallel using the SIMD-Single Instruction Multiple Data principle. 2) A superposition of states for input, internal and output variables, which makes it possible to compactly represent data and combine computational processes and phenomena in space and time. 3) Reversibility of quantum gates, arrays and operators, which makes it possible to significantly simplify the analysis of digital circuits in the test generation and fault detection.

II. DESIGN AND TEST AS THE XOR-RELATIONSHIP

Relation is a structure of interrelated components that determines the properties of a process or phenomenon in time or space. The structure determines the properties of the components, process or phenomenon, but not vice versa. Relation-signature is primary, media components are secondary. The alphabet is the carrier of a relation defined by operations (signatures) on symbols. Relations are decisive in creating effective mathematical theories, data structures, algorithms, architectures, models, methods, technologies, materials, services, software and hardware applications, cyber-physical and social systems, including economics, healthcare, transportation, law and order, ecology and statehood. The power of the relationship, as an integral set and the quality of mutual

relationships between the components, forms a metric that makes it possible to identify the effectiveness of the structure.

The technological space model associated with the Design and Test is also determined by (xor-) relations between the three structure components (Functionality, Test, List of faults), which form the transitive closure of a triangle: $F \oplus T \oplus L = 0$ that is essentially characteristic test equation. The transitive closure relation is the most effective among all the types of interaction, since it allows determining the third component, having any two components [8].

Procedures for test synthesis, fault simulation, and fault detection can be reduced to the xor interaction on the graph (Fig. 5) of the four nodes-components, where U represents the actual computing device. Here F and U create an image and a preimage, therefore they have a one-to-one correspondence for all essential components.

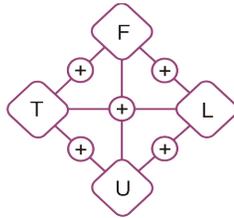


Fig. 5. The interaction graph of the test model components

The interaction graph gives rise to four basic triangles, which form 12 theoretically possible Design and Test tasks:

$T \oplus F \oplus L = 0$	$T \oplus L \oplus U = 0$	$T \oplus F \oplus U = 0$	$F \oplus L \oplus U = 0$
1) $T = F \oplus L$	4) $T = L \oplus U$	7) $T = F \oplus U$	10) $F = L \oplus U$
2) $F = T \oplus L$	5) $L = T \oplus U$	8) $F = T \oplus U$	11) $L = F \oplus U$
3) $L = T \oplus F$	6) $U = T \oplus L$	9) $U = T \oplus F$	12) $U = F \oplus L$

The introduction of an additional node U into the interaction graph (F, U, T, L) of the components Design and Test extends the functionality of the model of xor relations, including the carrying out of real experiments. The table columns define the following classes of Design and Test tasks:

Class 1 – theoretical experiments on the model of functionality: 1) Test generation based on the model for a given fault list. 2) Building a model of functionality based on a given test and fault list. 3) Fault simulation of functionality in a given test.

Class 2 – real experiments on the device: 4) Test generation through physical emulating faults of the device. 5) Determining the list of detected faults of a device on a given test. 6) Diagnosing or determining the real device in the presence of a test and faults.

Class 3 – test experiments without faults: 7) Test generation by comparing the results of the analysis of the model and the real device. 8) Synthesis of a reference model of functionality by testing a real device on a given test. 9) Diagnosing a real device with a test and a reference model.

Class 4 - online testing in operation: 10) Determining the functionality of the model during the online operation of the real device with the specified faults. 11) Online fault diagnosis of the

device in operation by comparison with the reference model. 12) Diagnosing a real device in the presence of faults and a reference model.

The presented testing space of digital devices forms all possible methods for the synthesis and analysis of data, including classic and most common technologies, which are characterized by the following equations: 1, 3, 5, 8, 9. The above equations can be combined into groups of methods for defining: tests, functionality models, fault list and real device:

- 1) $T = F \oplus L$; 4) $F = T \oplus L$; 7) $L = T \oplus F$; 10) $U = T \oplus L$;
- 2) $T = U \oplus L$; 5) $F = U \oplus L$; 8) $L = T \oplus U$; 11) $U = T \oplus F$;
- 3) $T = F \oplus U$; 6) $F = T \oplus U$; 9) $L = F \oplus U$; 12) $U = F \oplus L$.

All constructions represented in relations is characterized by the property of transitive reversibility of each triad of relations on a full graph, when it is possible to define the third component by any two components. At the same time, the representation format of each component in the equation should be identical in terms of the metrics of the parameters and the dimensions of the vectors or matrices. Further, we consider methods for testing and diagnosing digital systems in the metric of qubit vectors [8] describing the behavior of functional elements.

III. QUBIT DATA STRUCTURES FOR THE DIAGNOSIS

As an example illustrating the advantages of quantum (qubit) diagnosis of states, we consider the Deutsch algorithm [9], which solves the problem of recognizing the Boolean functions $f(X) = \{0, 1, x, \text{not}x\}$ of one variable. To distinguish the functions $f(X) = \{0, x\}$ and $f(X) = \{\text{not}x, 1\}$, the classical algorithm needs one measurement on the test pattern $X=0$, which is illustrated by the following transformation of truth tables for four primitive functions:

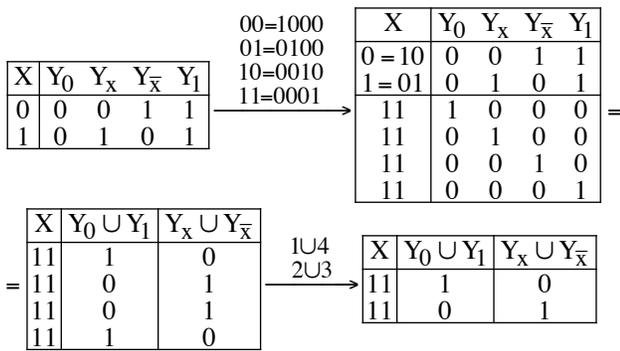
X	Y_0	X	\bar{X}	Y_1	$\xrightarrow{x=\{0,1\}}$	X	$Y_0 \cup X$	$\bar{X} \cup Y_1$
0	0	0	1	1		0	0	1
1	0	1	0	1		1	x	x

However, to distinguish the functions $f(X) = \{0, 1\}$ and $f(X) = \{x, \text{not}x\}$, the classical algorithm requires two measurements in two time frames ($X_1=0, X_2=1$), since it is considered that pairwise minimization of the columns $\{0, 1\}$ and $\{x, \text{not}x\}$ by uniting them for primitive functions is impossible:

$$\begin{array}{l} X_1 = 0 \rightarrow Y = 0 \\ X_2 = 1 \rightarrow Y = 0 \end{array} \rightarrow Y_0;$$

$$\begin{array}{l} X_1 = 0 \rightarrow Y = 1 \\ X_2 = 1 \rightarrow Y = 1 \end{array} \rightarrow Y_1.$$

It is established that for a quantum algorithm, one measurement is enough. This is followed by a simple proof of the fact that on a classical computer, it is possible to get the possibility of uniting the functions mentioned above, as well as the “quantum” result of diagnosis based on a single measurement. To be able to unite the mentioned pairs of functions, it is necessary to define the input values (0=10, 1=01), and also the qubit codes of output values in the unitary coding format:



The simplest transformations of rows of a table of unitarily coded functions by superposing (uniting) column vectors, followed by uniting pairwise identical rows (1 and 4; 2 and 3) make it possible to get a compact representation of two truth tables:

X	Y ₀ ∪ Y ₁	Y _x ∪ Y _{x̄}
11	1	0
11	0	1

The advantage of the obtained table lies in the distinguishability of two states (columns) for one formal measurement on the integrally written one-line test, using the unitary coding of the input and output rows of truth tables. The unitary code is attractive by the parallelism of computations due to the ability to minimize or unite in time and space any combination of state codes of primitive functions. This also applies to quantum computing, which is not a “marvelous” tool in comparison with the theoretical possibilities of classical computing. Any quantum algorithms are implemented in the classical version by increasing the memory for the unitary representation of primitive alphabets, models, functions and data structures. The use of quantum data structures and algorithms allows classical computing to increase the speed of solving practical problems by increasing hardware costs.

Next, the analysis of unitary matrices describing the functionality of primitive quantum elements is performed in order to the subsequent use of the mentioned above data structures for parallel diagnostics of logical functions.

The table of all logical functions of two variables can be represented as a test that recognizes 16 states of a digital device having 4 output lines:

x1x2	Y0	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13	Y14	Y15
00	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
01	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
10	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
11	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

The equation of recognizing the technical state by the reaction R(T) of the device on the full test T is as follows: $D = R(T) \oplus S_i = 0$. For regular structures, the diagnosis of the technical state is determined by the address of the corresponding column: $D = S[R(T)]$. Here the response of the device to the full test is the address of the column forming the technical state of the digital device.

If the value of only one output is known, which is zero or one, then the power of the set of identifiable (classified) states is always equal to half of all functions or $(1/2)(2^{**n})$. Otherwise,

one measurement on the test can identify at least half of the specified states. A Deutsch example can be reduced to recognizing columns with numbers Y6 and Y9 by performing xor operations on them: $0110 \oplus 1001 = 1111$. Any measurement – a test vector of the specified set: 00, 01, 10, 11 recognizes the states 0110 and 1001 apart.

In the general case, the set of recognizable states on an arbitrary test is the set of states belonging to the test response formed by calculating a list based on the following expression: $D = D \cup S_i \leftarrow R(T) \wedge S_i = S_i$. This procedure is parallel when making conjunction of a column and a ternary $(0, 1, X=\{0,1\})$ vector-response, but it is sequential when processing all columns of the function table.

The following table represents a two-stroke alphabet describing the automaton transitions of digital devices encoded in the unitary code. Such coding provides a unique opportunity to recognize arbitrary two different sets of symbol-states in the space of one variable $X=\{00,01,10,11\}$ using one measurement. The same does the quantum Deutsch algorithm.

x1x2	J	H	B	E	I	P	C	Q	S	O	F	A	L	V	Y
Q=00	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
E=01	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1
H=10	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1
J=11	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

The scaling of the functional space of two variables is associated with unitary coding of already 16 states, arranged in a matrix of dimension $16 \times 2^{**16}$, where all possible combinations consisting of 16 values will be presented. The result of diagnosis on such qubit unitary encoded matrix structure will be identical: two arbitrary column-states are recognized using one test measurement.

Special cases of the Deutsch-algorithm are further considered. Any two equal symmetric subsets of functions of n variables are always recognized in one measurement. For example, for two symmetric sets: $S = \{00, 11\}$ and $P = \{01, 10\}$, the operation of symmetric difference or the intersection $S \oplus P = \emptyset$ in the integral metric of the two-stroke cubic calculus [8] is equal to the empty set. A similar result is obtained if the mentioned symbols are unitary encoded: $S = \{00, 11\} = \{1000, 0001\}$ and $P = \{01, 10\} = \{0100, 0010\}$, operation of symmetric difference or intersection of superpositions of symbols-states $S \wedge P = 1001 \wedge 0110 = 0000$ or $S \oplus P = 1001 \oplus 0110 = 1111$. This means that these function subsets are recognized using any one measurement on any test pattern. The xor-interaction (comparison) of a pair of asymmetric functions and, or (and-not, or-not) gives a symmetric function xor: $(0001) \text{ xor } (0111) = (0110)$, $(1110) \text{ xor } (1000) = (0110)$.

Thus, in the positional code it is not always possible to minimize the logical functions of n variables. This means that in such a space it is impossible to define two arbitrary subsets by two vectors obtained as a result of minimization. The transition to the functional space of unitary codes successfully solves this problem, where an arbitrary function is always representable in an explicit form by two always non-intersecting vectors. This property is proof of recognition of two arbitrary subsets of states using a single test measurement. The price for such properties is the substantial initial redundancy of data structures for converting truth tables into unitary codes or qubit coverages of logical elements [8].

IV. DIAGNOSIS OF LOGICAL FUNCTIONS USING QUBIT DATA STRUCTURES

The proposed quantum method for fault detection based on qubit data structures [8, 10] uses parallel logical (set-theoretic) operations on rows of the table of detected faults. At that, the logical operations of uniting and intersecting the zero and unit rows of the table, corresponding to the positive and negative test results of the digital device forming the binary reactions of the observed outputs, are performed:

$$F = \left(\bigcup_{\forall R_i=1} Q_{ij} \right) \setminus \left(\bigcup_{\forall R_i=0} Q_{ij} \right) = \left(\bigvee_{\exists R_i=1} Q_{ij} \right) \wedge \left(\overline{\bigvee_{\exists R_i=0} Q_{ij}} \right).$$

The data structures are represented by the fault table on the Cartesian product of test patterns and the set of variables of the object under diagnosis, where each cell is two bits: the first of them identifies the detected stuck-at-0 (10) and the second one – stuck-at-1 (01):

$$\begin{aligned} Q &= \{F, T, L\}, \\ Q &= Q_{ij}, i = \overline{1, m}; j = \overline{1, n}; \\ F &= (F_1, F_2, \dots, F_j, \dots, F_n), \\ F_j &= \{10 \equiv 0; 01 \equiv 1; 11 = \{0, 1\}; 00 = \emptyset\}; \\ T &= (T_1, T_2, \dots, T_i, \dots, T_m); \\ L &= (L_1, L_2, \dots, L_i, \dots, L_n). \end{aligned}$$

The superposition of faults (two units on one line-cell) makes it possible to significantly minimize data structures for storing information in order to subsequent fault detection when performing a diagnostic experiment online.

To test the qubit fault detection method, the logic circuit shown in Fig. 6 is further proposed, which has 6 and-not elements, 11 lines, 5 inputs and two outputs.

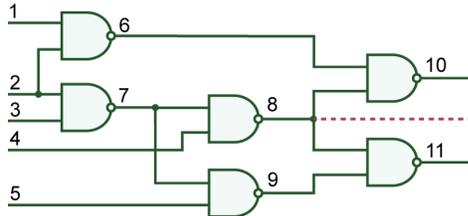


Fig. 6. ISCAS-circuit for verification

The following unitarily coded table illustrates the diagnostic process by uniting fault set, which form invalid output states on test patterns {T1-R10; T5-R11; T6-R10, R11; T8-R11}:

Q = Q _{ij}	1	2	3	4	5	6	7	8	9	10	11	R ₁₀	R ₁₁
T ₁	01	10	01	00	10	00	10	10	00	10	01	1	0
T ₂	10	00	10	00	01	10	00	00	10	01	10	0	0
T ₃	00	01	01	00	00	01	10	01	01	10	10	0	0
T ₄	10	00	01	00	10	00	01	00	10	01	01	0	0
T ₅	00	10	00	01	00	01	00	10	00	10	10	0	1
T ₆	01	10	00	00	10	00	00	01	10	01	10	1	1
T ₇	01	00	00	10	00	00	01	00	10	01	01	0	0
T ₈	00	10	10	01	01	10	00	00	00	01	10	0	1
Q ₁	01	11	11	01	11	11	10	11	10	11	11	1	1
Q ₀	11	01	11	10	11	11	11	01	11	11	11	0	0
F	00	10	00	01	00	00	00	10	00	00	00	1/0	1/0

The disjunction of the rows T1, T5, T6, T8 forms the vector Q1, which collects all possible faults detected on test patterns. The vector Q0 unites all impossible and undetectable faults on test patterns using the rows T2, T3, T4, T7. Subtraction of all impossible faults from all possible ones gives the result in the form of three faults, encoded as F2 = 10; F4 = 01; F8 = 10. Thus, execution of two register or-operations in parallel allows determining three possible faults, each of which can occur in the logic circuit:

$$F = \{2^0, 4^1, 8^0\}.$$

A condition of existence a single stuck-at fault in the logical circuit is more stringent, the use of which leads to the definition of faults based on the following expression:

$$F = \left(\bigcap_{\forall R_i=1} Q_{ij} \right) \setminus \left(\bigcup_{\forall R_i=0} Q_{ij} \right) = \left(\bigwedge_{\exists R_i=1} Q_{ij} \right) \wedge \left(\overline{\bigvee_{\exists R_i=0} Q_{ij}} \right).$$

Application of the formula clarifies the result of the diagnosis and reduces it to the form: $F = \{2^0\}$ due to the inconsistency of fault codes by and-operation in columns 4 and 8. The condition of the presence in the logic circuit of a single stuck-at fault puts into focus the following statement.

Statement. If the 00 or 01 coordinate exists in the fault table column, which creates incorrectness R=1 on the observed outputs associated with fault 10 on the remaining coordinates of the column, such a single fault (10) is impossible in the logic circuit.

Proof. Suppose that on n test patterns there is a discrepancy in the external outputs of the reference and real values of the signals. In this case, the n-1 coordinate of the column under consideration has the value 10 (01) and only one n-coordinate has the value 01 (10). If we assume, that there is a fault 10 in the logic circuit, then there must also be a fault 10 on n-coordinate, which defines an incorrect state of the outputs. But under the terms of simulation such a fault is absent there. Therefore, it is impossible to assume that there is a fault 10 in the circuit. This is also confirmed by the formal result – the empty intersection of all column coordinates associated with incorrect states of the circuit outputs:

$$F = \left(\bigwedge_{\exists R_i=1} Q_{ij} \right) = \begin{cases} 10 \wedge 10 \wedge \dots \wedge 10 \wedge 01 = 00; \\ 10 \wedge 10 \wedge \dots \wedge 10 \wedge 00 = 00. \end{cases}$$

This also applies to n-coordinate state, which is identified by a signal of the empty set 00, interaction with which also makes impossible the presence in the logic circuit of a single stuck-at-0 (code 10).

Thus, the combination of unitary coding of stuck-at faults allows reducing the dimension of the fault detection table to the number of lines in the circuit, to perform parallel logical operations on the rows of the table in order to significantly improve the performance of the method.

V. MEMORY-DRIVEN COMPUTING ARCHITECTURES

Modification of memory-driven computing consists in inserting the control unit into memory (Fig. 7, left), which makes it possible to eliminate the heterogeneity of the architecture and

bring it to pure memory, which implements control, computing and data storage modules before and after processing.

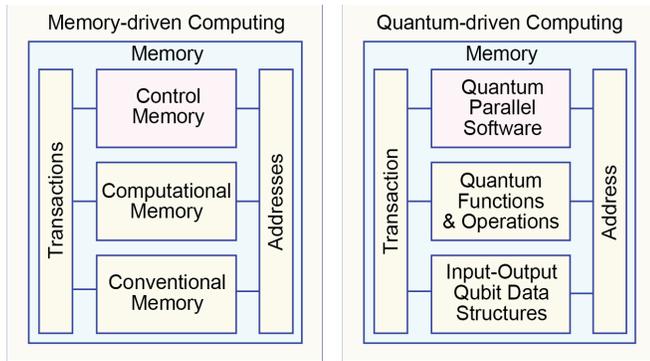


Fig. 7. Architectures: memory-driven and quantum computing

The structure is aimed at improving performance in solving combinatorial problems due to the high parallelism of quantum algorithms based on the use of qubit data structures. Memory-driven computing contains Computational Memory, which is represented by coverages of functional primitives, implementing logical (arithmetic) operations. Control Memory is a computational process control algorithm that reads data from Conventional Memory, processes them using Computational Memory, and writes the results to regular memory for storage.

Such a memory computing structure is a typical one that can be extended to solving a practical problem related to digital circuit simulation based on qubit data structures, Fig. 8, left. The presented memory architecture contains the following components: structure-algorithm, function-operations, input-output data. In particular, a memory-driven SIMD (Single Instruction Multiple Data) computing architecture is focused on parallel data processing, typical for quantum-driven digital circuit analysis using the characteristic equation $M=Q[M(X)]$.

Here, the control memory block is represented by the above equation; computational memory – qubit coverages, which form functional primitives; conventional memory – input and output data before and after processing. In other words, the creation of data structures focused on Quantum Computing means encoding states and processes in unitary code, making it possible to apply parallel logical operations to sets of data of the same type. Thus, the architecture for the ALU-free computational process using only memory, on which address transactions are implemented, forms MemComputing, the structural version of it is MAT-Computing (Memory-Address-Transaction). Another example of memory computing is shown in Fig. 8, right, which is focused on parallel data processing in order to significantly improve performance in solving problems of pattern, text, image and figure recognition.

Thus, combining quantum computing and the architecture of computational memory provides science and practice with a new technological culture of quantum memory-driven computing, which integrally gains the advantages of structural homogeneity and parallelism of processing big data through eliminating transactions between memory and the ALU processor, and also eliminating quantum operations of superposition and entanglement.

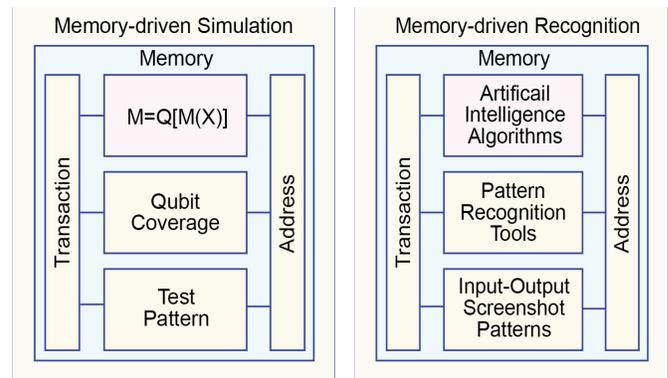


Fig. 8. Memory-driven quantum computing: Simulation and Recognition

VI. CONCLUSION

The effectiveness of the proposed models and methods for the synthesis and analysis of digital systems is metrically evaluated by performance parameters, memory hardware costs, as well as the quality of the designed processes and products. The scientific novelty of the proposed research consists in combining quantum computing technologies, qubit data structures, unitary coding, parallel execution of logical operations and modern methods for the synthesis and analysis of digital systems, which allows increasing the speed of synthesis and analysis of Design and Test processes by an order of magnitude: 1) memory-driven technology for designing, testing and diagnosing digital systems-on-chips is proposed, which differs from the classical theory by the absence of traditional logic for the synthesis and analysis of computing devices and the use of qubit data structures of quantum computing [11, 12].

2) A model of a universal digital primitive is presented that excludes a truth table and is represented by a qubit binary vector of unitary state coding, which adequately describes the behavior of any functionality, combinational or sequential type.

3) Methods for the synthesis of functionalities based on qubit vectors stored in memory, which compactly describe the behavior of computer components, are considered. Parallel methods for analyzing digital architectures based on qubit vectors are proposed, which allow effectively simulate the behavior of computing devices by using only read-write transactions in memory, and also to perform logical operations in parallel to recognize technical states.

4) Methods for testing and diagnosing digital systems based on qubit vector structures are considered, which make it possible to effectively generate tests and simulate faults of computing devices by using parallel vector logic operations.

REFERENCES

- [1] G. Benenti, G. Casati, G. Strini, "Principles of Quantum Computation and Information," vol. 1: Basic Concepts, World Scientific, 2004.
- [2] M.A. Nielsen, I.L. Chuang, "Quantum Computation and Quantum Information," Cambridge University Press, 2010.
- [3] M. Nakahara, "Quantum computing: an overview," Mathematical Aspects of Quantum Computing, 2007.
- [4] M.G. Whitney, "Practical Fault Tolerance for Quantum Circuits," A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Computer Science in the Graduate Division of the University of California, Berkeley, 2009.

- [5] K.M. Svore, B.M. Terhal, D.P. DiVincenzo, "Local Fault-Tolerant Quantum Computation," [Online]. Available: <http://research.microsoft.com/pubs/143764/local2005.pdf>
- [6] O. Golubitsky, D. Maslov, "A Study of Optimal 4-bit Reversible Toffoli Circuits and Their Synthesis," IEEE Transactions on Computers, 2011, pp. 1-14.
- [7] J.P. Hayes, I. Polian, B. Becker, "Testing for Missing-Gate Faults in Reversible Circuits," Proc. Asian Test Symposium, Taiwan, November 2004.
- [8] V. Hahanov, "Cyber Physical Computing for IoT-driven Services," Springer, New York, 2018.
- [9] R. J. Lipton, K. W. Regan, "Quantum Algorithms via Linear Algebra," MIT Press eBook, 2014.
- [10] V. Hahanov, W. Gharibi, S. Chumachenko, E. Litvinova, I. Iemelianov, M. Liubarskyi, "Quantum Data Structures for SoC Component Testing," International Journal of Design, Analysis & Tools for Integrated Circuits & Systems, Oct. 2017, vol. 6, iss. 1, P. 23.
- [11] V. Hahanov, E. Litvinova, S. Chumachenko, I. Iemelianov, M. Liubarskyi, "Qubit test synthesis for the black box functionalities," Proc. of 5th Prague Embedded Systems Workshop, June 29-30, 2017, Roztoky u Prahy, Czech Republic, pp.45-51.
- [12] Hahanov V. Quantum Sequencer for the Minimal Test Synthesis of Black-box Functionality / V. Hahanov, S. Chumachenko, I. Hahanova, I. Iemelianov, I. Hahanov // Proc. of IEEE East-West Design and Test Symposium.– Novi Sad.– October, 2017.– P.445-450.

The Fault Tolerant CMOS Logical C-Element for Digital Devices Resistant to Single Nuclear Particles

Yuri V. Katunin

Department of Analog and Digital Blocks Design
Scientific Research Institute of System Analysis, Russian Academy of Sciences
Moscow, Russia
katunin@cs.niisi.ras.ru

Vladimir Ya. Stenin

Department of Electronics
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Moscow, Russia
vystenin@mephi.ru

Abstract—The TCAD simulation results of the new CMOS logical C-element based on the trigger with reduced switching delay and two tristate inverters designed using 65-nm bulk CMOS technology are presented. Transistors of the element are divided into two groups so that charge collection from the track of a single nuclear particle by transistors of one group only cannot cause the C-element trigger to fail. Charge collection from tracks with linear energy transfer of 60 MeV·cm²/mg does not lead to changes of the logical function of the element and to failures when the C-element transmits common-mode logic signals. The nature of charge collection from tracks does not significantly depend on operation mode of the C-element as well as on the moment of setting the common-mode signals for state switching or antiphase signals for state storage.

Keywords—65-nm bulk CMOS technology, charge collection, C-element, fault tolerance, non-stationary state, nuclear particle, single event transient, TCAD simulation

I. INTRODUCTION

A logical C-element is an element with two inputs. In the case of common-mode input signals the C-element transmits their logical state. If the input signals are antiphase the C-element stores the last common-mode state of the inputs. The C-element was proposed [1] as part of asynchronous logic. It was developed in methods for designing CMOS elements that are resistant to single effects of nuclear particles under the names such as keeper-less C-element [1], SERT - single-event resistant topology [2], and tri-state inverter transmission gate [3]. The D trigger with increased noise immunity [4, 5] is based on the spacing of transistors into two groups (Spaced Transistor Groups DICE - STG DICE) so that charge collection from the track of single nuclear particle by only one of the transistor groups does not result in the trigger failure. In this work, we propose the new C-element developed using the STG DICE methodology. The aim of the work is to increase the fault tolerance of the C-element without significant performance costs and to obtain reliable quantitative evaluation of its fault tolerance during the virtual experimental study using TCAD.

II. THE HARDENED C-ELEMENT

The scheme of the C-element is shown in Fig. 1. The C-element consists of two tri-state inverters TRInv 1 and TRInv 2 and the STG DICE trigger realized as two groups of transistors

Group 1 or Group 2. Transistors of groups form the ring of the four elementary triggers with four nodes ABCD in which the state “0” (ABCD = 0101) or the state “1” (ABCD = 1010) can be written.

The first group of STG DICE D trigger (Group 1) contains $N_D P_A$ and $N_A P_B P_{B1} P_{B2}$ transistors. The second group (Group 2) contains transistors $N_B N_{B1} N_{B2} P_C$ and $N_C P_D$. Inverters TRInv 1 and TRInv 2 with the high-impedance state contain transistors $N_{1.1} N_{1.2} P_{1.1} P_{1.2}$ and $N_{2.1} N_{2.2} P_{2.1} P_{2.2}$, respectively. Transistors N_{B1} , N_{B2} , P_{B1} , P_{B2} introduced into the circuit to reduce the switching delay of the STG DICE trigger.

Simulations of CMOS structures on 65-nm bulk technology (with a channel length of 65 nm) were carried out using 3-D TCAD transistor models presented in the work [6]. The 3-D device model of the C-element is shown in Fig. 2. The

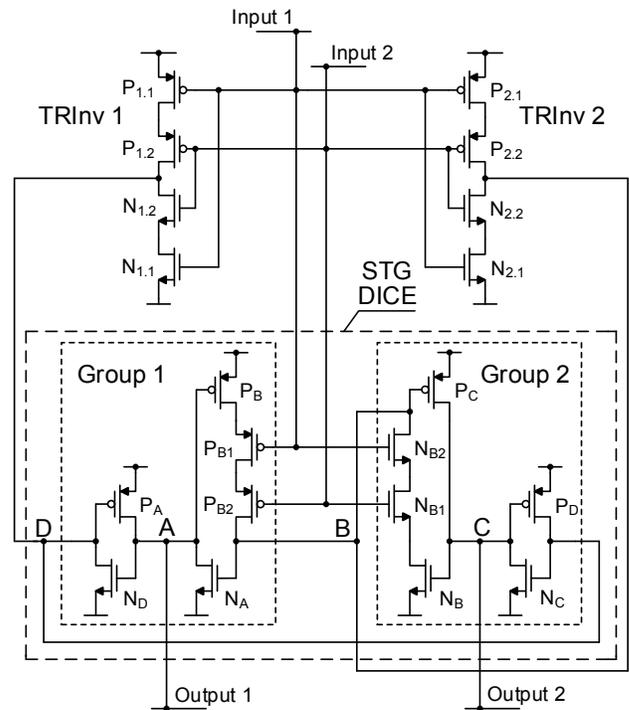


Fig. 1. The scheme of the C-element based on the STG DICE trigger and the inverters TRInv 1, TRInv 2.

The work was done in the framework of the State assignment, project 0065-2019-0008.

regions of a shallow trench isolation covering the silicon regions of transistors to the depth of 400 nm are hidden.

The total dimensions of the 3-D device structure, including areas not occupied by transistors, are $6.4 \mu\text{m} \times 10.9 \mu\text{m}$. The substrate thickness is $3.0 \mu\text{m}$. The distance between the drain centers of N_D and N_B transistors from different groups is $2.35 \mu\text{m}$. Between the regions of NMOS and PMOS transistors of logic elements in the device structure (Fig. 2) there are heavily doped n^+ and p^+ regions, which are elements of guard rings. The substrate chip doped by boron has the concentration of 10^{16}cm^{-3} . The region doped with boron using Gaussian profile has the peak concentration of $5 \times 10^{18} \text{cm}^{-3}$ at the depth of $1.25 \mu\text{m}$. The doping zone is of $\pm 0.4 \mu\text{m}$. Device layers have Gaussian doping profile with the peak concentration of $2 \times 10^{18} \text{cm}^{-3}$ at the depths of $0.65 \mu\text{m}$ with boron for NMOS transistors and arsenic in the n-well for PMOS transistors.

Charge collection by transistors from tracks passing in a silicon crystal parallel to the surface at the depth of 100 nm was adopted as the test impact in the work. The amount of charge generated on the track depends on the energy transfer by the particle on the track, while the energy component of a charge generation characterized by the linear energy transfer (LET) by the particle to the track [7]. The results were obtained during simulation in Sentaurus Device at the temperature of 25°C and the supply voltage of 1.0 V for CMOS structure designed using the 65 nm CMOS bulk technology for particles with $\text{LET} = 60 \text{MeV}\cdot\text{cm}^2/\text{mg}$.

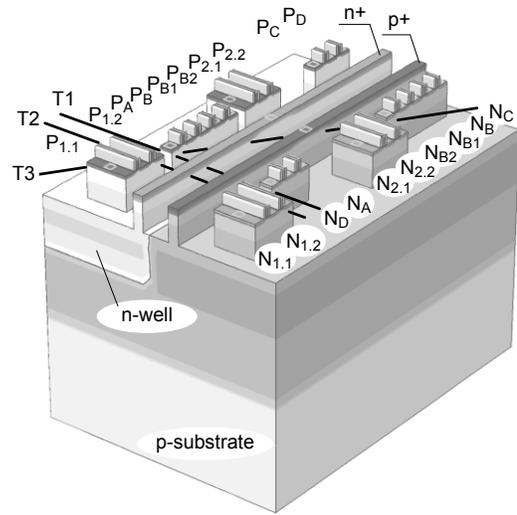
III. SWITCHING MODE OF THE LOGIC STATE OF THE TRIGGER BY THE COMMON-MODE INPUT SIGNALS

The main modes of operation of the C-element: 1) switching the logical state of the trigger with two input common-mode signals; 2) transition of the input signal "0" or "1" to the output of the trigger and C-element; 3) storage of the previous state of the trigger, when the input signals become antiphase.

A. Charge Collection from the Track T1 Passing through Transistors P_A and N_D of the Group 1

The time dependences of the voltage at the nodes of the C-element are shown in Fig. 3; the initial logic state of the trigger is "1" (the state of the nodes ABCD is 1010). The track T1 of a particle with $\text{LET} = 60 \text{MeV}\cdot\text{cm}^2/\text{mg}$ passes in the device layer at the depth of 100 nm under the drain regions of the transistors P_A and N_D of the Group 1. Simultaneous change of input signals from the state "1" to "0", leading to the trigger switching, occurs at $t_{\text{SWT}} = 500 \text{ps}$. The track T1 of a particle is formed 105 ps before the start of switching at $t_1 = 395 \text{ps}$ (Fig. 3a) and 10 ps after the start of switching at $t_2 = 510 \text{ps}$ (Fig. 3b).

Before switching the trigger from state "1" (Fig. 3a) to state "0" at $t \leq t_{\text{SWT}} = 500 \text{ps}$ the transistors $N_{1.1}$ and $N_{1.2}$ (Fig. 1) of the inverter TRInv 1 are open, the transistors $P_{1.1}$ and $P_{1.2}$ are closed, and the outputs of both inverters set the voltage $V_B = V_D = 0$ at the nodes B and D. The transistors P_A and N_D are in active mode. The reverse-biased drain pn-junction of the open transistor P_A collects the charge from the track. At the beginning of the charge collection at $t = 400 \text{ps}$ the transistor P_A goes to inverse mode with increasing voltage at the node A up to 1.75V (Fig. 3a). The open transistor N_D also goes into inverse mode with a voltage at the node D (at the drain of N_D) $V_D = -0.2 \text{V}$



Simulation of charge collection from the track T1 with the start of collecting 10 ps after the start of trigger switching from state "1" to "0" showed (Fig. 3b) that the nature of voltage changes at the nodes A and D has the same character as in fig. 3a when charge was collected from the same track for the trigger state "0". Moreover, the nodes C and D retain the initial levels until the moment of switching at 630 ps, and the switching of the trigger is completed 122 ps after changing the input signals of the C-element. Thus, the nature of charge collection doesn't significantly depend on the moment when the common-mode input signals switching occurs. Durations of non-stationary states, including the switching delays, are given in the Table.

B. Charge Collection from the Track T2 Passing through Transistors $P_{1,2}$ and $N_{1,2}$ of TRInv 1

The time dependences of the voltage at the nodes of the C-element are shown in Fig. 4; the initial logic state of the trigger is "1" (ABCD = 1010). The track T2 of a particle with LET = 60 MeV·cm²/mg passes in the device layer at the depth of 100 nm under the drain regions of the transistors $P_{1,2}$ and $N_{1,2}$ of the inverter TRInv 1. Simultaneous change of input signals from the state "1" to "0", leading to the trigger switching, occurs at $t_{\text{SWT}} = 500$ ps. The track T1 of a particle is formed 20 ps after the start of switching at $t_1 = 520$ ps (Fig. 4a) and 100 ps after the start of switching at $t_2 = 600$ ps (Fig. 4b).

In the initial state before switching, when the common-mode input signals $V_{\text{IN}1} = V_{\text{IN}2} = 1$ V, the transistors $P_{1,2}$ and $P_{1,1}$ of the inverter TRInv 1 are closed. After changing the input signals to the levels $V_{\text{IN}1} = V_{\text{IN}2} = 0$ these transistors become opened, and transistors $N_{1,2}$ and $N_{1,1}$ become closed. Thus, after switching, the charge (electrons) from the track T2 is collected by the drain pn-junction of the transistor $N_{1,2}$. After 20 ps (Fig. 4a) or 100 ps (Fig. 4b) from the start of trigger switching, when charge collection by the transistor $N_{1,2}$ begins, the voltage at the drain of the transistor $N_{1,2}$ (at the node D) decreases rapidly from 1 V to 0.3 V and then to 0.1 V.

In this case, the voltages at the nodes A, B, C practically correspond to the levels that should be after switching the trigger of the C-element (Fig. 4a and Fig. 4b). The switching duration (switching delay) of the C-element trigger is 20-35 ps, the duration of the non-stationary state of the trigger is 250 ps (Fig. 4a) and 217 ps (Fig. 4b) and practically does not depend on the time interval between the start of charge collection from the track and the start of switching the C-element.

C. Storage Mode of Logical State "1", Followed by the Transmission mode of "0" when Collecting the Charge from the Tracks T1 or T2

The storage mode (when antiphase signals are applied to the inputs) with subsequent transition to the mode of transmission of common-mode input signals to the output of the C-element is characterized by switching the state of the trigger described in the Section III.A. The only difference is that the state "1" of the trigger (ABCD = 1010) at the initial stage of time up to 500 ps is saved due to transition of both inverters TRInv 1 and TRInv 2 in the high-impedance state after setting the antiphase input signals $V_{\text{IN}1} = 1$ V and $V_{\text{IN}2} = 0$.

The dependences of voltage at the trigger nodes in the storage mode, when the charge is collected from the track T1

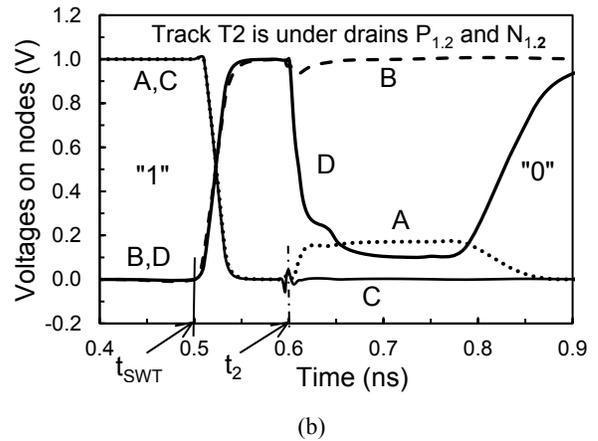
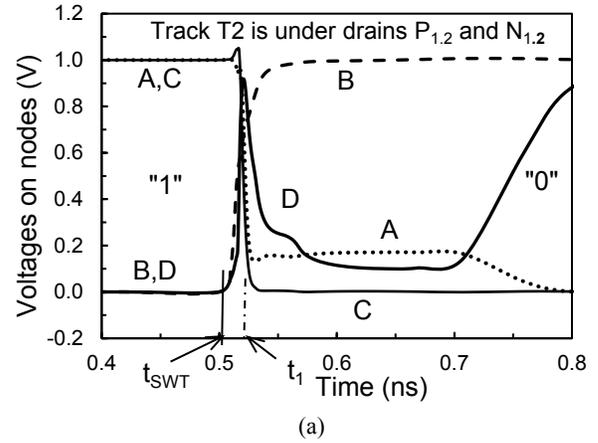


Fig. 4. The voltages at the nodes of the C-element; the track T2 is under the drains of the transistors $P_{1,2}$ and $N_{1,2}$; the trigger switches from "1" to "0" at $t_{\text{SWT}} = 500$ ps; the beginning of the charge collection: (a) at the moment $t_1 = 520$ ps, (b) at the moment $t_2 = 600$ ps.

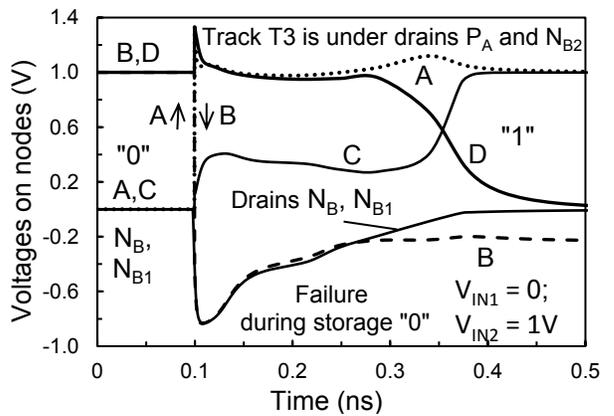


Fig. 5. The voltages at the nodes of the C-element; the track T3 is under the drains of the transistors of the two group (the transistor P_A in Group 1 and the transistor P_{B2} in Group 2) in the storage state with $V_{\text{IN}1} = 0$, $V_{\text{IN}2} = 1$ V; the initial state of the trigger is "0"; the beginning of the charge collection at $t_1 = 100$ ps.

passing through the transistors P_A and N_D , practically do not differ from the dependencies shown in Fig. 3 for "Switching from "1" to "0" mode. Similarly, the dependences of voltage at the trigger nodes in the storage mode with the transition to the mode of transmission of common-mode input signals to the

TABLE. DURATIONS OF NON-STATIONARY STATES (IN PS) DURING THE CHARGE COLLECTION FROM THE TRACKS T1 AND T2 AT LINEAR ENERGY TRANSFER BY PARTICLE TO TRACK 60 MEV·CM²/MG

Track	The Moment of the Impact	Switching from “1” to “0”	Storage “1” with transition to transmission “0”
T1	100 ps before switching	123	120
T1	20 ps after switching	122	122
T1	100 ps after switching	93	96
T2	100 ps before switching	251	253
T2	20 ps after switching	249	249
T2	100 ps after switching	217	217

output, when the charge is collected from the track T2 passing through the transistors $P_{1,2}$ and $N_{1,2}$, also do not differ from the dependences shown in Fig. 4. Duration values of non-stationary states caused by the charge collection from the track T1 or track T2 are given in the Table for the mode “Storage “1” with transition to transmission “0”. Thus, the nature of the charge collection from the track does not significantly depend on the fact that the trigger was previously in the storage mode.

D. Storage Mode of Logical State “0” when Collecting the Charge from the Track T3

Fig. 5 shows the voltage dependences at the nodes of the C-element trigger when the charge is collected from the track T3 in the storage mode; the voltages at the inputs of the element are $V_{IN1} = 0$, $V_{IN2} = 1$ V and the initial stored state of the trigger is logical zero “0”. In this mode the transistor $N_{2,2}$ of the inverter TRInv 2 is open because $V_{IN2} = 1$ V, but transistors N_B , N_{B1} , N_{B2} , $N_{2,1}$ located near are closed (see Fig. 2). These transistors are the main charge collectors from the track T3.

At the beginning of the charge collection each of the transistors N_B , N_{B1} , N_{B2} , $N_{2,1}$, $N_{2,2}$ goes into inverse offset mode and the trigger has the same voltages at the nodes as in the mode of signal transmission when $V_{IN1} = V_{IN2} = 0$. The fail of the storing state “0” is due to the fact that the inverters TRInv 1 and TRInv 2 go to a high-impedance state when voltages at the inputs become $V_{IN1} = 0$, $V_{IN2} = 1$ V. These inverters do not participate in the restoration of the initial states of the trigger nodes B and D after the end of the non-stationary state of the trigger caused by the charge collection from the track T3. The mode with the initial state of the trigger “1” is characterized by the fact that the state “1” is saved and there is no failure. Durations of the non-stationary states caused by the charge collection from the track T3, are 250-300 ps.

IV. ANALYSIS OF SIMULATION RESULTS

At the common-mode switching of the input signals of the C-element always after the non-stationary state, caused by the charge collection from the track, the trigger transits to the state, which is setting by the common-mode signals at the inputs. This behavior of the trigger does not depend on the moment of beginning the charge collection from the track (before switching, at the time of switching the input common-mode signals, or after switching the signals).

During the transmission of common-mode input signals “0” or “1” to the trigger output, the charge collection from the track

leads to a temporary non-stationary state at the trigger nodes, after which the states of the nodes are restored.

After the storage mode, when the C-element input signals were antiphase, the transition common-mode signals puts the trigger nodes in the corresponding logical state with the delay less 25-30 ps.

V. CONCLUSION

The proposed CMOS C-element on the STG DICE cell is the basis for design of fault-tolerant asynchronous logic circuits. Charge collection by transistors of only one from two groups of the STG DICE trigger till the LET 60 MeV·cm²/mg does not cause failures of the C-element when transmitting common-mode logic signals.

In the switching by common-mode signals or in the mode of transmission the common-mode signal “0” or “1”, regardless of the charge collection beginning, after the end of the single event transition the C-element goes in the state corresponding to the input signal. In the mode of the state storage, when the input signals of the C-element is antiphase, the trigger fail may occur. In this case, the transition to the common-mode signals returns the output of the trigger to the state corresponding to common-mode signals with a delay of less than 25-30 ps.

The C-element on the STG DICE cell can be useful in fault-tolerant developments at the 28–65 nm CMOS technology nodes.

REFERENCES

- [1] D.E. Muller, and W.S. Bartky, “A theory of asynchronous circuits,” Proc. of International Symposium on the theory of switching, Cambridge, M.A.: Harvard University Press, 1959, pp. 204–243.
- [2] J. Gambles, K. Hass, and S. Whitaker, “Radiation-hardness of ultra-low power CMOS VLSI,” Proc. of 11th NASA Symposium on VLSI Design, 2003, pp. 1–6.
- [3] R.J. Baker, CMOS Circuit Design, Layout, and Simulation (IEEE Press Series on Microelectronic Systems). – Hoboken, New Jersey: John Wiley & Sons, Inc., 2010, p. 351.
- [4] V.Ya. Stenin, Yu.V. Katunin, and P.V. Stepanov, “Upset-Resilient RAM on STG DICE Memory Elements with the Spaced Transistors into Two Groups,” Russian. Microelectronics, vol. 45, no. 6, 2016, pp. 419–432.
- [5] Yu.V. Katunin, and V.Ya. Stenin, “Simulation of single event effects in STG DICE memory cells,” Russian Microelectronics, vol. 47, no. 1, 2018, pp. 20–33.
- [6] R. Garg, S.P. Khatri, Analysis and design of resilient VLSI circuits: mitigating soft errors and process variations. New York: Springer, 2010. pp. 194–205.
- [7] Soft errors in modern electronic systems / Editor M. Nicolaidis. New York: Springer, 2011. pp. 35–37.

Development of a Simulation Tool to Estimate Electricity Consumption and Determine the Optimum Cooling System for Data Centers

Beyzanur Toprak

Department of Computer Engineering
Faculty of Technology, Selcuk University
Konya, Turkey
beyza.toprak58@gmail.com

Beyza Nur Bora

Department of Computer Engineering
Faculty of Technology, Selcuk University
Konya, Turkey
beyzaabora@hotmail.com

Gül Nihal Güğül, Member IEEE

Department of Computer Engineering
Faculty of Technology, Selcuk University
Konya, Turkey
gul.gugul@selcuk.edu.tr

Abstract—Electricity consumption of datacenters is increasing significantly with the growth in Information Technology sector. Electricity consumption of datacenters occupies approximately 2% of total electricity consumption in the USA which makes IT Sector second most electricity-consuming sector. Cooling system in a traditional datacenter is responsible of 30-60% of final electricity consumption. In this study, a software is developed to estimate the final energy demand of a datacenter and electricity consumption of the datacenter for variable cooling systems by using Visual Studio C# interface and SQL server. Climate data, physical properties of the datacenter building and IT equipment situated in datacenter inputs are used in order to software to calculate annual energy demand of datacenter. Then electricity consumption of the datacenter is presented on monthly basis and electricity demand of datacenter for each cooling system is listed on annual basis. Developed software is validated by using power demand value, physical properties of datacenter building information and equipment list of an existing data center. Typical climate data insertion tool and economic analyses will be added to the software in order to be used worldwide. Therefore, this is an ongoing study.

Keywords—datacenter, cooling system, energy efficiency, SQL, c#

I. INTRODUCTION

Data centers consist of cooling systems and cabinets that include servers, switch devices and storage arrays. In a standard datacenter approximately half of the electricity is consumed in cabinets and the remaining is consumed for cooling system [1]. Increase in requirement of data storage space and server performance resulted in significant increase in cooling demand respectively. Datacenters must be kept at acceptable temperatures to perform their functions. A significant percent of old datacenters are still being cooled by floor standing cooling systems. However, cooling energy demand in energy efficient datacenters can be lowered as much as 10% of the total energy consumption, compared to 50% in the air-cooled datacenters [2]. Water cooling is the second common option for cooling which is newer and energy efficient than air cooling. Free cooling is one more and most efficient option in which environmental temperatures are used such as lakes, sea and cold outside air.

Energy demand of a datacenter should be estimated before construction in order to analyze energy saving potential for different scenarios. It is estimated that the typical data center today could hold up to 30% more IT equipment with existing power demand and cooling system capacity if the capacity was properly managed. Today data centers are not able to fully utilize its available cooling capacity, which increases electricity consumption by 20% or more compared to a properly managed system. Therefore, capacity management tools results in better utilization of power and cooling resources and reduce power demand. Capacity management is the ability to quantify the supply and the demand for power and cooling [3]. There are many datacenter capacity management tools developed for management of existing datacenters ([4], [5], [6]) by monitoring IT equipment's and cooling system in real time. These software tools are developed in order to analyze and present the monitored energy consumption of devices and cooling system in an existing datacenter. A simulation environment for cloud computing data centers is developed in a study which is designed to capture details of the energy consumed by data center components (servers, switches, and links). The simulation results demonstrate the effectiveness of the simulator in utilizing power management schema [7]. Another study is conducted to presents a coordinated cooling and load management strategy that is based on a holistic and modular approach to data center modeling that represents explicitly the coupling between the thermal and computational subsystems. Simulation results for a small example illustrate the potential for a coordinated control strategy to achieve better energy management than traditional schemes that control the computational and cooling subsystems separately [8]. These tools are developed to manage datacenters after construction.

In addition to capacity management tools, many studies are conducted to calculate the energy saving potential in datacenters before construction however few of them developed software for these calculations. A study is conducted in China to develop software defined network (SDN) technique and explore a new solution to energy-aware flow scheduling, such as scheduling flows in the time dimension and using exclusive routing for each flow. In this study simulations and testbed experiments showed that exclusive routing can effectively save network energy

compared to regular fair-sharing routing (FSR) [9]. A study is conducted by University of Michigan and Schneider Electric to develop a model for total data center power. In this study, power models of each data center component are collected from variable sources. Component models were suitable for integration into a detailed data center simulator. In this study the design for such a simulator is outlined [10]. Another study is conducted in USA by Syracuse University and IBM to develop and experimentally validate a model that will allow data center designers to evaluate the energy saving potential of various loop configurations and operating strategies. The modeling methodology of this study couples a thermodynamic model of the cooling equipment and a hydraulic pipe network for hydraulic characterization. This model has the ability to capture off-design operating conditions of the data center cooling system caused by changes in ambient conditions and fluctuations in power demand. The model is validated based on an existing IT data center and chiller plant located in Poughkeepsie, New York [11]. SIMOPEK project is developed in Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities to develop methods and software components for modeling, simulating, and optimizing the cooling infrastructure of a data center. The models take into account both the highly dynamic load behavior of the computing system as well as new technological components and concepts for recycling the generated waste heat. Developed tool provides a virtual reconfiguration of the cooling circuits before physically rebuilding the system with the goal to efficiently use and re-use energy [12].

As it is seen from the review of previous studies and to the authors' best knowledge, a study in order to develop a modeling and simulation software in which location, building's physical properties and user selected IT equipment of datacenter are taken into account simultaneously to calculate the electricity demand of datacenter for different cooling systems is not conducted. The aim of this study is to develop a software in order to estimate energy saving potential of datacenters for different locations, physical properties of building, IT equipment and indoor temperature set values before construction of a datacenter or retrofit of an existing datacenter and calculate electricity demand of the datacenter for floor standing air conditioner and free cooling system.

II. METHODOLOGY

In this study a software tool was developed in visual studio c# interface. In developed software firstly climate data of the location of datacenter is selected. Then physical model of the datacenter building is developed by using TS 825 Heat Insulation Regulation in Buildings Standard of Turkey [13]. After modeling the datacenter building, datacenter equipment list is selected by software user or energy demand per unit area is inserted if already known. Then annual energy consumption of the datacenter equipment is calculated by using the physical features of datacenter building, climate data of the region and the electricity consumption information per unit area of the all cabinets in datacenter. Finally, annual cooling demand of datacenter is calculated for different cooling systems. Flow chart of the simulation tool is given in Fig. 1.

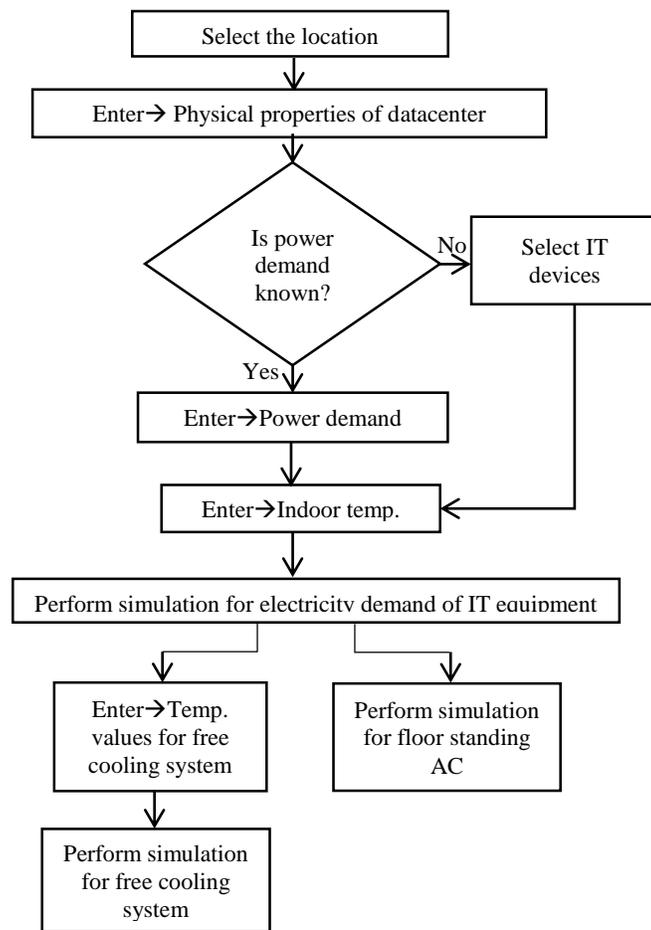


Fig. 1 Flow chart of the simulation tool

A. Climate Database

Climate database is developed in SQL interface. General Directorate of Meteorology of Turkey shares monthly average values of approximately 80-90 years of dry bulb temperature, solar radiation and sunshine duration [14]. In this study climate data of four biggest cities of Turkey that are Ankara, Istanbul, Izmir and Konya are used. However, to extend the user profile of the software hourly epw (EnergyPlus weather) data is going to be used in software. Epw weather data includes data of more than 2100 locations worldwide and already available for Ankara, Istanbul and Izmir.

B. Modeling Datacenter Building

In order to calculate cooling demand, firstly location of the datacenter is selected by user and climate data in database is used according to selected location. Then user has to provide size and U (Thermal Conductivity Coefficient) values of each exterior component of the building to the software. Then specific heat loss, heat loss through conduction and convection of the building are calculated by using the equations in Table 1.

TABLE I. EQUATIONS USED IN CALCULATION OF COOLING DEMAND

Description of Equation	Equation	Abbreviations in Equation
Specific heat loss (H) of the building	$H = H_c + H_v$ (1)	H_v : Heat loss by ventilation H_c : Heat loss by conduction and convection
Heat loss through conduction and convection	$HT = \sum(A \times U) + l \times UI$ (2) $\sum(A \times U) = U_{ow}A_{ow} + U_wA_w + U_{od}A_{od} + 0.8 \times U_cA_c + 0.5 \times U_bA_b + U_{ba}A_{ba} + 0.5 \times U_{lt}A_{lt}$ (3)	U_{ow} : Thermal conductivity (TC) coefficient of outer wall, W/m^2K U_w : Window TC coefficient, W/m^2K U_{od} : TC coefficient of outer door, W/m^2K , U_c : TC coefficient of the ceiling, W/m^2K , U_b : TC coefficient of basement, W/m^2K , U_{ba} : TC coefficient of the base in contact with the outdoor air, W/m^2K , U_{lt} : TC coefficient of building elements in contact with indoor environments at low temperatures, W/m^2K , A_{ow} : Area of outer wall, m^2 A_w : Window area, m^2 A_{od} : The area of the outer door, m^2 A_c : Ceiling area, m^2 A_b : Basement area sitting on the floor, m^2 A_{ba} : Basement area in contact with outdoor, m^2 A_{lt} : The area of the building elements in contact with the indoor environments at low temperatures, m^2 l : Thermal bridge length, m U_l : Linear permeability of the thermal bridge, W/mK
Monthly cooling demand	$CD = (P \times A \times 0.6 + H \times (\Theta_i - \Theta_o)) \times \frac{24 \times 30}{1000}$ (4)	CD : Cooling Demand, kWh/month H : Heat Loss, $W/^\circ C$ Θ_i : Indoor Temperature, $^\circ C$ Θ_o : Outdoor Temperature, $^\circ C$ P : Power consumption of datacenter, W/m^2 A : Basement area of datacenter, m^2

C. Calculation of Electricity Consumption of IT Equipment

After the details of physical properties of the datacenter building are provided, power demand per unit area of IT equipment in datacenter is calculated by using the information of selected devices from device database of the software. Power demands of equipment in database are obtained from datasheets of each device. Generally typical power consumption of IT equipment are nearly 60% percent of maximum power given in datasheets [15]. Therefore, power demand of each device is multiplied by number of device and power demand of all selected devices are summed. Then sum of power demand is multiplied by 0.6 in order to obtain typical power consumption value. Finally, power demand is divided to basement area of datacenter to obtain power demand value per unit area (W/m^2). In case of availability of electricity consumption of IT equipment per unit area, user is also allowed to enter the available value instead of selecting all devices separately.

D. Calculation of Electricity Demand of the Datacenter for Different Cooling systems

Cooling demand of the datacenter is calculated by using equation (4) in Table 1 and information provided by user in previous forms of software. Then by using final demand for

cooling, electricity demand of the datacenter for free cooling and floor standing air conditioner is calculated.

- Free cooling

Free cooling electricity demand depends on number of hours smaller than indoor set temperature. Therefore, free cooling electricity demand is calculated on hourly basis. However, in this project due to the availability of monthly weather data, electricity demand is calculated on monthly basis. Running modes of a free cooling system are free cooling mode, partial free cooling mode and mechanical cooling mode. Running modes of system according to temperature are given in Table 4.

TABLE II. OPERATING TEMPERATURES FOR COOLING MODES

Cooling Mode	Operating Temperature ($^\circ C$)
Free cooling	$T_{ot} \leq T_{st}$
Partial free cooling	$T_{st} < T_{ot} < T_{rt}$
Mechanical cooling	$T_{ot} \geq T_{rt}$

T_{ot} is outside air temperature, T_{st} is supply air temperature and T_{rt} is return air temperature in Table 2.

The Power Usage Effectiveness (PUE) shows the electricity efficiency of data centers [16] which is the ratio of total

electricity consumption in a datacenter to the electricity consumed by the IT equipment. In order to calculate PUE value, electricity consumption of cabinets and cooling system should be known. In a study conducted in Ankara, Turkey electricity consumption of IT equipment in a datacenter and free cooling system in datacenter are measured [17]. In the study, average electricity consumption of cabinets is found as 274,248 kWh/year (31,306 W) [17] and electricity consumption of free cooling system is measured as given in Table 5.

TABLE III. ELECTRICITY CONSUMPTION OF FREE COOLING SYSTEM

Mode	Power (W), [17]	Ratio of Power of Cooling system to Power of IT equipment
Free Cooling Mode	1709	0.0545
Mechanical Mode	39444	1.2599
Pump Power	1512	0.0482

In Table 3, in addition to the measured free cooling system modes, ratio of power of each free cooling system mode to power of IT equipment is given due to the parallel change of electricity consumption of free cooling system to electricity consumption of IT equipment.

In this study free cooling electricity consumption of datacenter is calculated by multiplying the power demand of IT equipment with the ratios in Table 3 for each mode and free cooling mode of the cooling system is determined according to comparison of indoor set temperature provided by software user and outdoor temperature in weather database of the location of datacenter.

- *Air cooling*

In this study cooling demand of datacenter is calculated by using equation (4) in Table 1. Traditional air-cooled datacenters often operate at a Power Usage Effectiveness rating of about 1.6 [18] which means cooling add 60 % increase on the power of IT equipment. Power demand of IT equipment is multiplied by 0.6 in equation (4) in order to calculate heat dissipation from IT equipment. Therefore, cooling demand of datacenter is equal to 60% of power demand of IT equipment in datacenter when outdoor temperature is lower than internal temperature. Cooling demand of the datacenter is equal to 60% of power demand of datacenter and heat gain from sun when outdoor temperature is higher than internal temperature.

E. Validation of Software

After the development of software, an existing datacenter located in Konya, Turkey is modelled in the software in order to validate the software. Then energy demand of the modelled datacenter is estimated by the software and obtained result is compared with the instant power value of the existing datacenter. Physical properties of the existing datacenter are given in Table 4. In addition to physical properties, list of datacenter equipment in datacenter are given in Table 5.

TABLE IV. PHYSICAL PROPERTIES OF THE DATACENTER

Physical property	Value
Dimensions, meters	4×4×2.5
Doors, meters	2.2×1.2
Windows, meters	1.5×1
Indoor temperature, °C	21
Cooling equipment	4 pieces of 45,000 Btu floor standing air conditioner [19]

TABLE V. DATACENTER EQUIPMENT IN EXISTING DATACENTER

Equipment	Piece	Properties of equipment
Switch	5	48 Port Switch 5*7 A
Switch	1	48 port Switch 9 A
Switch	1	80 Port Switch 16 A
Switch	1	72 port switch 34 A
Firewall	1	17/20 Gbps firewall 24 6 A
Storage	1	125x 2.5-inch Storage 13 A
Server	5	2-socket, 2U rack server 5*8 A
Server	3	Xeon Proc server 3*7 A
Server	3	900 Mhz server 3*6 A
Controller	2	WLAN Mobility controller 2*4 A

III. RESULTS AND DISCUSSION

In this section screenshots of developed software are given with the simulation results of existing datacenter of which’s detailed information are given in Table 4 and Table 5.

There are five sub-forms in software which are “Location”, “Building”, “Electricity Demand”, “Simulate” and “Cooling Equipment”. In first sub-form user is requested to select the location of datacenter shown in Fig. 2.

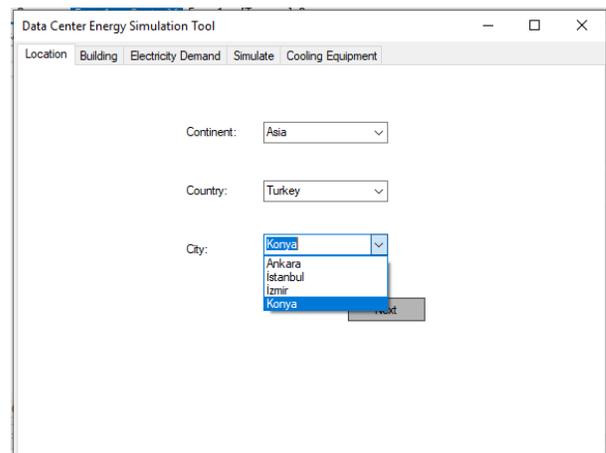


Fig. 2: Location sub-form of software

Next form allows user to provide physical properties of datacenter building as shown in Fig. 3.

Fig. 3: Building sub-form of software

Fig. 5: Simulate sub-form

Then user is requested to enter power demand per unit area of the datacenter if known. If power demand is not known, user can also enter equipment list of datacenter and in this case power demand is calculated by software. “Electricity Demand” sub-form is shown in Fig. 4.

Fig. 4: Electricity Demand sub-form

As shown in Fig. 4 power demand of the IT equipment list given in Table 5 in existing datacenter is estimated as 717 W/m² by the developed software. According to the information provided by datacenter management, instant current driven by the datacenter equipment is measured as 44 amperes by datacenter staff which is 605 W/m² (44 A×220 Volt=9680 W=605 W/m²) which is in nearly 84% accuracy with estimated value. Power demand of datacenter equipment is calculated by datasheets of equipment in software and due to the change of power demand parallel to data traffic the estimation of software is assumed to be in acceptable range.

After the estimation of power demand of IT equipment, user is allowed to simulate the software. In “Simulation” form user has to enter indoor set temperature of the datacenter and then monthly cooling demand of the datacenter is calculated and given as shown in Fig 5.

Indoor set temperature of existing datacenter is 21 °C. Only July and August is higher than 21°C in monthly average temperature data of Konya. As a result of this, as seen in Fig 5, all months except July and August, cooling demand is equal to 60 % percent of electricity consumption of IT equipment. On July and August, cooling demand is sum of heat gain from sun and IT equipment. According to Table 4 existing datacenter is cooled by 4 pieces of floor standing air conditioners. Four floor standing air conditioners measured to draw approximately 10×4=40 ampere current by datacenter staff that is 6336 kWh/month (40×220×24×30/1000=6336). This measurement is conducted in winter months when outdoor temperature is lower than indoor set temperature and is estimated as 4955 kWh/month by developed software as seen in Fig 5 which is in nearly 80% accuracy.

Finally, electricity consumption of the datacenter for different free cooling system and floor standing air conditioner are calculated in “cooling equipment” sub-form as shown in Fig 6 and Fig. 7. For free cooling system user has to provide supply temperature and return temperature of the water circulated in datacenter room.

Fig. 6 Cooling Equipment sub-form for free cooling

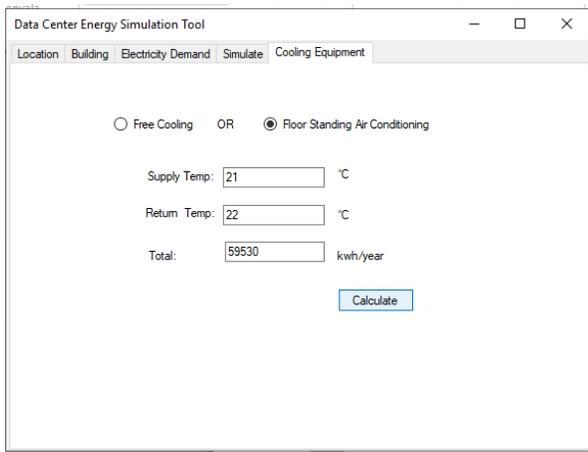


Fig. 7 Floor standing air conditioning sub-forms of software

As seen from Fig 6 and Fig 7, free cooling system is calculated to consume 75% less electricity compared to floor standing cooling system in Konya climate. Cooling system of existing datacenter is measured to draw approximately $10 \times 4 = 40$ ampere current by datacenter staff that is 6336 kWh/month ($40 \times 220 \times 24 \times 30 / 1000 = 6336$).

A. Calibration of Software

Energy demand of an existing datacenter is estimated by developed software. Final current drawn by datacenter is measured as 84 amperes which leads to nearly 18.5 kW ($84 \text{ A} \times 220 \text{ Volt} = 18,480 \text{ W}$) instantaneous power demand and 13,305 kWh monthly electricity consumption that is assumed to be constant. Estimated and measured annual electricity demand of the datacenter cooling system, IT equipment and final electricity consumptions are given in Table 6. Estimated and measured final electricity consumptions are found as 158,765 kWh/year and 159,667 kWh/year respectively that are in acceptable proximity.

TABLE VI. ESTIMATED AND MEASURED ENERGY DEMAND IN EXISTING DATACENTER

	Cooling, kWh/year	IT Equipment, kWh/year	Final Energy, kWh/year
Estimated	59,600	99,165	158,765
Measured	76,032	83,635	159,667
Proximity	78%	84%	99%

IV. CONCLUSION

In this study, a software is developed to estimate final energy demand and determine the optimum cooling system for a data center in order to minimize energy demand. Developed software allows a user to estimate electricity consumption of

cooling system of a datacenter for different locations, building properties, IT equipment and indoor set temperature. By using developed software, user can determine the properties of optimum datacenter before selecting the location, constructing the building, buying equipment and cooling system. Developed software is validated by modeling final energy demand and cooling demand of an existing data center of which's electricity demand is measured instantly. Estimated and measured electricity demand of the existing datacenter is found as 158,765 kWh/year and 159,667 kWh/year which are in acceptable proximity.

V. REFERENCES

- [1] C. R. Gil, "Energy Efficiency in Data Processing Centers", International Conference on Renewable Energies and Power Quality, Granada, 2010.
- [2] J. B. Marcinichen, J. A. Olivier and J. R. Thome, "On-chip two-phase cooling of datacenters: Cooling system and energy recovery evaluation", Applied Thermal Engineering, 41, pp. 36-51, 2012.
- [3] N. Rasmussen, "Power and Cooling Capacity Management for Data Centers," White Paper 150, 2012.
- [4] Schneider Electric, "StruxureWare Data Center Operation," 2019. <https://www.schneider-electric.com/en/product-range/61867-struxureware-data-center-operation/>.
- [5] Sunbird, "Data Center Infrastructure Management," 2019. <https://www.sunbirdcim.com/what-dcim>.
- [6] Opendcim, "Opendcim Data Center Infrastructure Management," 2019. <https://opendcim.org/>.
- [7] D. Kliazovich, P. BouvrySamee, U. Khan, "GreenCloud: a packet-level simulator of energy-aware cloud computing data centers", The Journal of Supercomputing, 262-3, pp. 1263-1283, 2012.
- [8] Luca Parolini, Bruno Sinopoli, Bruce H. Krogh, "Reducing Data Center Energy Consumption via Coordinated Cooling and Load Management". Carnegie Mellon University Pittsburgh, PA
- [9] D. Li, Y. Shang and C. Chen, "Software defined green data center network with exclusive routing", IEEE Conference on Computer Communications, Toronto, 2014.
- [10] S. Pelley, D. Meisner, T. F. Wenisch, J. W. VanGilder, "Understanding and Abstracting Total Data Center Power," 2009.
- [11] D. W. Demetriou, H. E. Khalifa, M. Iyengar, R. R. Schmidt, "Development and experimental validation of a thermo-hydraulic model for data centers," HVAC&R Research, 17-4, pp. 540-555, 2011.
- [12] T. Wilde, T. Clees, H. Schwichtenberg, H. Shoukourian, D. Labrenz, I. Torgovitskaia, M. Schnell, N. Hornung, B. Klaaßen, E. L. Alvarez, "Towards energy-efficient data center infrastructure –a holistic approach based on software for modeling, simulation, and (re)configuration of the energy network", 3rd International Conference on ICT for Sustainability, Copenhagen, Denmark, September 7-9, 2015.
- [13] TS 825 "Heat Insulation Regulation in Buildings Standard", 2018. http://www1.mmo.org.tr/resimler/dosya_ekler/cf3e258fbd3eb7_ek.pdf
- [14] General Directorate of Meteorology of Turkey, 2018. <https://www.mgm.gov.tr/veridegerlendirme/il-ve-ilceler-istatistik.aspx?k=A&m=ANKARA>.
- [15] Huawei, S5720-SI. 2019. <https://e.huawei.com>
- [16] Cole, David. Data Center Energy Efficiency. Mission Critical Magazine. 2011. <https://www.missioncriticalmagazine.com>
- [17] G. N. Güğül, "Free Cooling Potential of Turkey for Datacenters", European Journal of Science and Technology, 14, pp. 17-22, 2018.
- [18] J. Dai, M. M. Ohadi, D. Das ve M. G. Pecht, «Optimum Cooling of Data Centers,» New York, Springer, 2014.
- [19] 45000 Btu Air conditioner, 2018. <https://e-fjthermaklima.com/fujithermaklima-salon-tipi-FTHFX45BAC>

Terms of Arrangement Reckoning Self-Checking Embedded Check Circuits Based on Boolean Complement up to Constant-Weight Code ‘1-out-of-3’

Dmitry Efanov,
DSc, Professor at “Automation, Remote Control and Communication on Railway Transport”,
Russian University of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

Valery Sapozhnikov,
DSc, professor at “Automation and Remote Control on Railways” Department,
Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
port.at.pgups@gmail.com

Vladimir Sapozhnikov,
DSc, professor at “Automation and Remote Control on Railways” Department,
Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
at.pgups@gmail.com

German Osadchy,
Technical Director of Scientific and Technical Center “Integrated Monitoring Systems” LLC,
St. Petersburg, Russia
osgerman@mail.ru

Dmitry Pivovarov,
PhD student at “Automation and Remote Control on Railways” Department, Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
pivovarov.d.v.spb@gmail.com

Abstract—Presented research is dedicated to the synthesis problem of self-checking embedded check circuits (concurrent error-detection systems) in accordance with Boolean Complement Method concerning constant-weight codes. Existing limits per components structures of concurrent error-detection (CED) systems based on example of constant-weight code ‘1-out-of-3’ (1/3-code) application are being analyzed by authors. It was demonstrated that apart from the testing performance regarding the segment of Boolean Complement together with test meter within check circuit, it is essential to obtain reliable self-checking system per site under control plus logical segment. Laid down conditions reckoning fulfillment of the complete self-checking system structure of the CED, are based on Boolean Complement Method via 1/3-code. We introduced the needed examples, which allow us to visualize the trouble of components testing performance with consideration concerning the application of Boolean Complement Method for the arranged cases of self-checking discrete systems.

Keywords—self-checking embedded check circuit, concurrent error-detection system, Boolean Complement Method, constant-weight code, code ‘1-out-of-3’, self-checking structures

I. INTRODUCTION

One of the most important task concerning synthesis of robust and safe discrete systems of automated management is proper methods implementation aimed at failures detection in time with immediate critical data presentation to self-checking embedded systems (CED systems) [1 – 5]. There are a large variety of methods of synthesis per sets of CED, started from conventional backing-up method up to application of constant-weight codes while structural schema choice based on Check Bits Calculation or via Boolean Complement Method [6 – 11].

On Fig. 1 we may see generalized structural chart of CED system, where, logical unit under control $F(x)$ is being equipped with a special schema of supervision as a part of the segment of logic control $G(x)$ with the totally self-checking

checker (TSC). Segment of logic control adjust the data per several audit function and the checker evaluate between each other definitions of working plus control functions with a next manage signal issue [12].

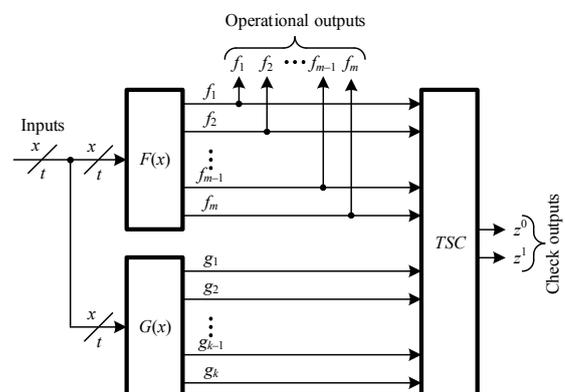


Fig. 1. Arranged structure of CED system.

For the arrangement of discrete devices with self-checked reckoning designated model of failures several terms should be completed. First of all, any logical device under control should obtain the required features for the proper testing performance, such as any failure out of designated matrix should be shown as an extortion of output data. Special check system is being implemented for the definition of the extortion of output data, where segment of logical supervision should be checkable one with TSC. The feature of self-control means the technical unit must be self-checking device with the entire safeguard from any failure of the designated matrix. The definition ‘characteristic of self-testing’ is considered the opportunity of each failure presentation via the coded combination per at least a single input port and as for the term of ‘protection from malfunction’, it means the impossibility of wrong

working coded variations installation in the event of any damage occurrence. Consequently, the synthesis procedure of self-checking systems of CED has several substantial limits.

For some scope of works, for instance, in [13 – 18], we can see the advantages of Boolean Complement Method in comparison with Check Bits Calculation one, related to better conditions of the CED arrangement plus synthesis possibility for simplest structures. In this case, maximum effectiveness while synthesis of the CED systems in the event of Boolean Complement Method is considered the constant-weight codes application with short length of applied words (r/n -codes, where r – weight of coded word, n – length of digital word). Hence, authors of presented paper did not pay enough attention reckoning the task of failure testing reliability within segments of main and control logic. Present work is covering the above case via example of research of the application possibility of 1/3-code during arrangement of the CED system in the event of single permanent embedded logical elements damages.

II. BASIC STRUCTURE OF BOOLEAN COMPLEMENT VIA 1/3-CODE

In case of supervision completion by means of logical devices of 1/3-code, those outputs to be divided by groups per three pieces in each one (we may apply equal outputs in different groups), and for each of such group self-circuit of control to be arranged and next outputs of the entire circuits of supervision should be unified within input of self-checked comparator being synthesized based on two-rail signals pressure modules of cascade connection [19]. On Fig. 2 basic structural scheme of CED system is being shown based on Boolean Complement Method of 1/3-code. The unique feature as compared with conventional structural schema of CED system is the segment of Boolean Complement, including the cascade of XOR elements. The above is essential for the alternation of working function vectors $\langle f_1 f_2 f_3 \rangle$ into coded words $\langle h_1 h_2 h_3 \rangle$, being the part of 1/3-code. Any vector of working function can be conversed into vector of 1/3-code with alteration of two digital order numbers f_i .

In the event of CED systems synthesis in accordance with structural schema being presented on Fig. 2, it is vital to ensure to supply total inputs $1/3-TSC$ with the entire digital words of 1/3-code, as well as to each element of composition per XOR obligatory tests, including combination {00; 01; 10; 11} [20]. The abovementioned is considered inputs supplement of CED sets via subset matrix of input combination.

Let us define which limits are the part of structural segments $F(x)$ and $G(x)$ within CED system being synthesized based on 1/3-code [13, 14, 21, 22], for the purpose of testability function insurance.

III. TERMS OF THE ENTIRE SELF-CHECKING FUNCTION ENSURANCE OF BASIC STRUCTURE

Within CED system there are four independent segments: segment of main logic $F(x)$, segment of control logic $G(x)$, segment of Boolean Complement Method with self-checking 1/3-TSC. Failure presence is designated per one of four segments only. Methods of total performance examination regarding the segment of Boolean Complement with tester are described in the abovementioned works. For the purpose of the total self-checking performance of basic structure arrangement we should ensure acquisition within the tester output any single failure event within segments structure $F(x)$ and $G(x)$.

The unique feature of 1/3-code considered whichever types of errors definition with any type of multiplicity within coded words, excluding double tampering, such as while simultaneous damage of on-bits with off-bits. This particular feature should be the matter of fact in the event of terms definition concerning the entire self-checking of basic structure per the structure of the CED system, consequently we should adjust such limits on segments $F(x)$ and $G(x)$, for the purpose that any single malfunction must not resulted in Boolean Complement segment failure being spotted on output via symmetrical error.

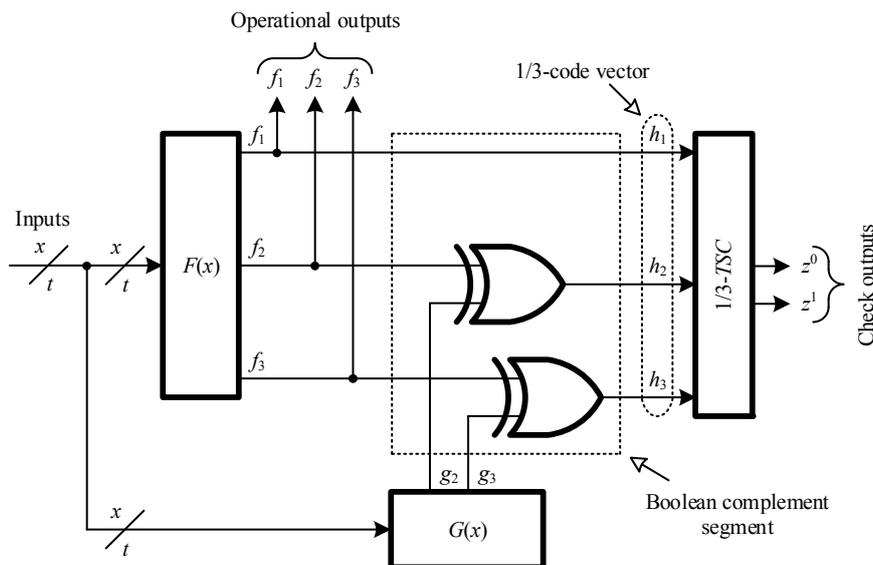


Fig. 2. Basic structural schema of the CED system.

Let us take a look at segment $F(x)$. In the event of malfunction insertion within segment on outputs f_1, f_2 and f_3 we may encounter with single, double or triple errors. In case of the single fault on the output of checker, the vector of $r=0$ or $r=2$ weight is being formed and this error is being recorded. In the event of the triple error on the output of checker, the vector of $r=2$ weight is being formed and this error is being recorded as well. There are two types of double errors: single direction (unidirectional) and multidirectional (symmetric)

ones. According to analysis we should emphasize that any of those errors may be monitored as well as it can be recorded on the output data of a tester meter. The abovementioned outcome depends on signal value of g_2 and g_3 , being worked out on the output of the segment $G(x)$. Let us call segment output $F(x)$ f_1 – passive, for the reason a signal is not being transformed, consequently outputs f_2 and f_3 – active.

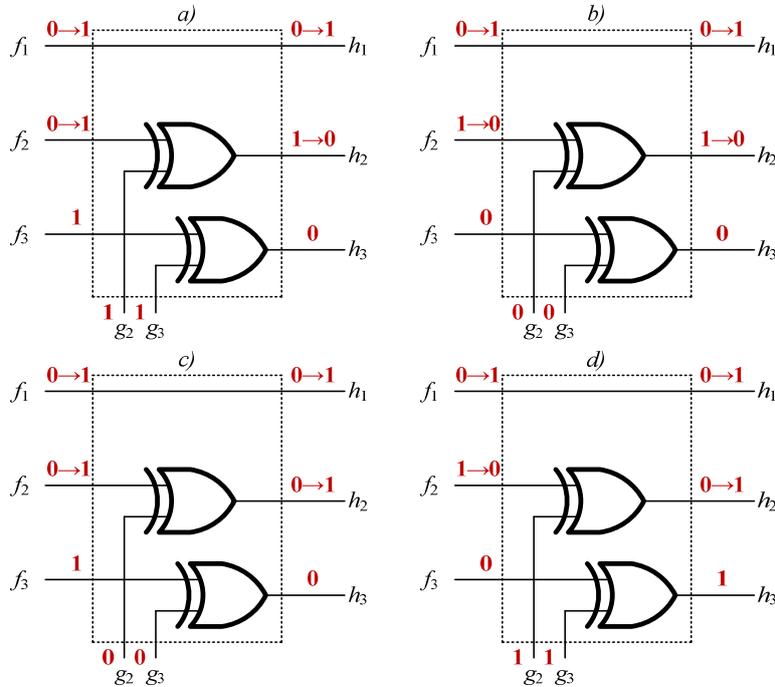


Fig. 3. Errors on the adjusted and uncontrolled outputs of the being supervised device.

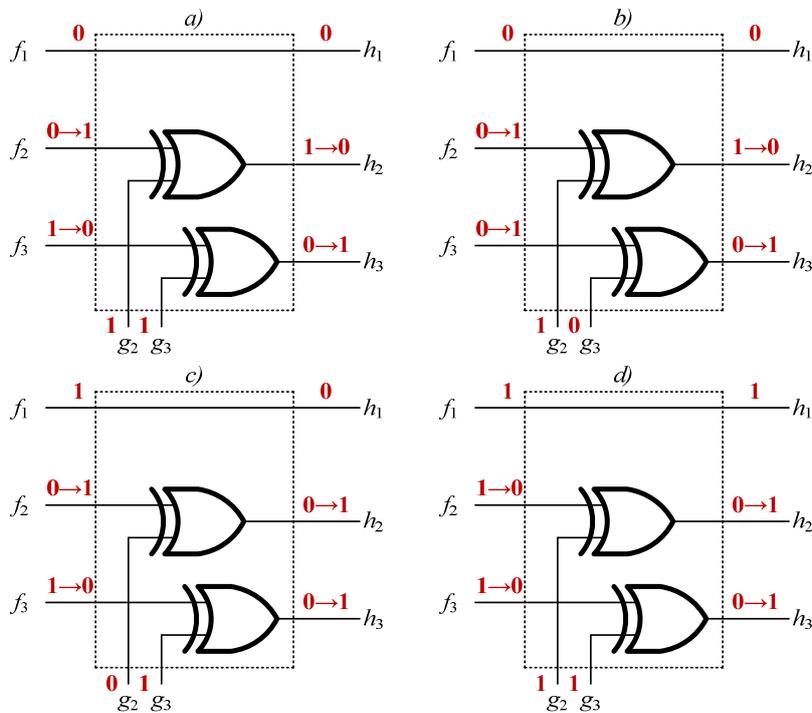


Fig. 4. Errors on the adjusted outputs of the being supervised device.

On Fig. 3 we can see the examples of basic structure performance for the option when errors are being shown simultaneously within active and passive outputs. Consequently, on Fig. 3, a), b) and c) were recorded single direction distortions per outputs f_1 and f_3 . Hence, in initial case on the checker inputs false vector was recorded $h_1h_2h_3=100$, so the mistake was not defined, as for the second alternative it was the vector 111 and that error was disclosed.

On Fig. 4 we may see the examples of basic structure performance for the condition when errors were being appeared on both active outputs at a time. Consequently, on Fig. 4 a) and b) multiple type distortions were recorded within outputs f_2 and f_3 . For the first case that error was defined, but for the second one it was just missed.

Correlation between signals value on segments outputs $F(x)$ and $G(x)$ should be distinct as follows:

Statement 1. Simultaneous warp of two values from three outputs of segment $F(x)$ should be detected within the following conditions:

a) In the event of signal distortion on passive f_1 and active f_b outputs ($b \in \{2,3\}$) the following terms should be fulfilled:

If $f_1 = f_b$, consequently $g_b = 0$, or

If $f_1 \neq f_b$, as a result $g_b = 1$.

b) In case of signal malfunction on active f_a and f_b ($a, b \in \{2,3\}$) the following conditions must be completed:

If $f_a \neq f_b$, thus $g_a \neq g_b$, or

If $f_a = f_b$, accordingly $g_a = g_b$.

As a fact, let us take a look, for example, on the initial episode. Assume that $g_b = 0$. Thereat $h_b = f_b \oplus g_b = f_b$. That is why in the event of $f_1 = f_b$ including the presence on errors on both tester's output we shall receive double single direction error.

If $g_b = 1$, consequently $h_b = f_b \oplus g_b = \overline{f_b}$. That is why, in the event of $f_1 \neq f_b$ plus double single direction error tester's output we shall get double single direction fault as well.

Let us suppose that recorded data per second option to be proved the same way. On Fig. 5 and Fig. 6 we may see several examples, which should clarify the appearance of non-detectable errors of Boolean Complement output.

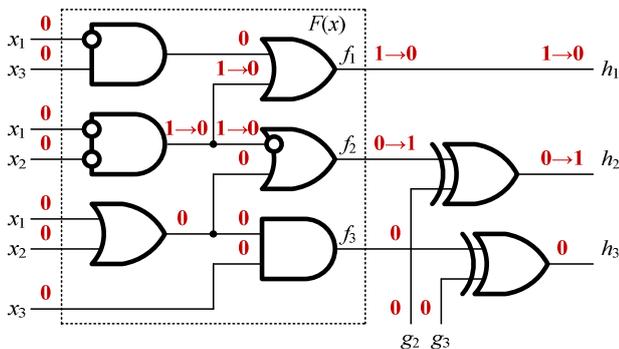


Fig. 5. Circuit of adjusted with non-adjustable outputs, where undetectable error may appear.

The condition in accordance with Statement 1, considered sophisticated for the supervision performance, so it can be used for the system of functional control arrangement with rather simple segments $F(x)$. Usually for multiple outputs schemas control is being arranged per groups of outputs. In this particular case, basic structure of 1/3-code may be applied per group out of three H^r -independent outputs monitoring upon the term of H^r -two outputs independence f_i and f_j (H^2 -independence) being presented by formula:

$$\frac{\partial f_i}{\partial y_t} \cdot \frac{\partial f_j}{\partial y_t} = 0, \quad i, j \in \{1,2,3\}, \quad (1)$$

Which should be correct per each logical element G_t , with function y_t realization on the way out.

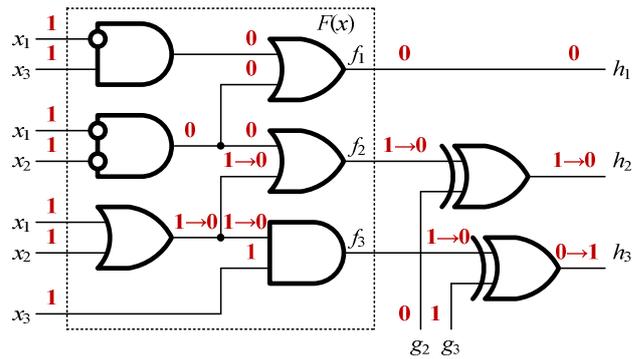


Fig. 6. Circuit of adjusted outputs where undetected error may appear.

Within group out of three H^r -independent outputs per each input set of the schema, the distortion is impossible more then per single output function. Meanwhile, the above trouble should be detected inside the basic structure! Hence, equal distortion should be spotted by simultaneous warping of the total three outputs, in this event (1) may be transformed as follows:

$$\frac{\partial f_i}{\partial y_t} \cdot \frac{\partial f_j}{\partial y_t} \cdot \overline{\frac{\partial f_p}{\partial y_t}} = 0, \quad i, j, p \in \{1,2,3\}. \quad (2)$$

Let us call the condition (2) as the H^3 -independence of three outputs.

Left part of the formula (2) should define the group of input sets, on which we may see the distortion factor for two out of three functions, excluding those sets of total three damaged functions.

On Fig. 7 we may see the circuit of three outputs. The entire pairs of outputs are out of independence status quo, but the above is absolutely in accordance with condition (2). That is the reason why the order of basic structure being presented in figure, show the total malfunction of logical elements of the main segment.

For arrangement of completely self-checkable structure for the circuit Fig. 7, we shall realize those circuits per the functions g_2 and g_3 separately (maintain the independence on both outputs of the logical control). In this case, we can ensure the influence of single error per one element only XOR and symmetrical error appearance must be impossible per outputs 1/3-TSC. We may compare the structure of present system with back-up kit. Meanwhile, it is possible, for instance, to

estimate degree of complexity reckoning its realization based on number of logical elements outputs, including the segment sophistication $F(x)$ equals $L_{F(x)}=14$, $L_{G(x)}=4$, $L_{XOR}=6$, $L_{TRC}=12$, $L_{1/3-TSC}=18$, $L_{NOT}=1$ [23].

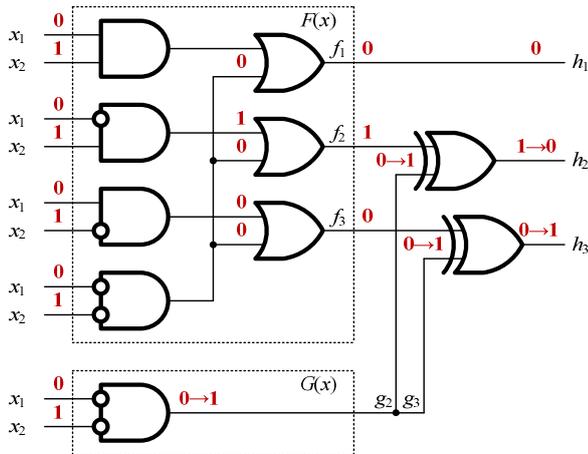


Fig. 7. Combinational circuit with three outputs.

For the CED system, shown on Fig. 7, we shall get

$$L_{CED} = L_{F(x)} + L_{G(x)} + 2L_{XOR} + L_{1/3-TSC} = 14 + 4 + 2 \cdot 6 + 18 = 48.$$

For the back-up system

$$L_D = 2L_{F(x)} + 3L_{NOT} + 2L_{TRC} = 2 \cdot 14 + 3 \cdot 1 + 2 \cdot 12 = 52.$$

Next let us take a look at the task of malfunction detection within the segment logical control $G(x)$, which has two outputs g_2 with g_3 only (see Fig. 2). Lone error for those outputs can be identified for the reason of single mistake evidence per outputs 1/3-TSC. Analogical to Statement 1 can be proved the next thesis.

Statement 2. Simultaneous outputs distortion on segment $G(x)$ can be identified on outputs if basic structure within following cases:

- a) If $g_a = f_b$ ($a, b \in \{2;3\}$), thus $f_a = f_b$,
- b) If $g_a \neq g_b$, consequently $f_a \neq f_b$.

On Fig. 8 we may see examples for the Statement 2 for the case, when signals g_2 and g_3 have the single direction distortion.

During arrangement of totally self-checked basic structure status of Statement 2 is the matter of control. Hence, for the reason the segment $G(x)$ is being synthesized separately from other segments of basic structure, it can be realized via the device H^2 -independent outputs. For example, within schema on Fig. 7 segment $G(x)$ is presented as a single element with function $x_1 x_2$. The error on the output of this element is $0 \rightarrow 1$ while receiving on input this set $x_1 x_2$ is not being recorded. Consequently, to get that entirely self-checking schema, present element should obtain the back-up unit by means of outputs g_2 and g_3 physical division.

In the event of creating totally self-checked multi outputs combined circuits H^3 -independent output groups may be applied as well as [24], unidirectional independent [25] with H^2 -independent groups [26]. On Fig. 9 we may see double level layout with six outputs, with absence of the abovementioned outputs group, but two sets H^3 -independent outputs $\{f_1, f_3, f_5\}$ and $\{f_2, f_4, f_6\}$. That is why we may create completely self-checking CED system of the present circuit without modification but via back-up either via two basic structures of 1/3-code.

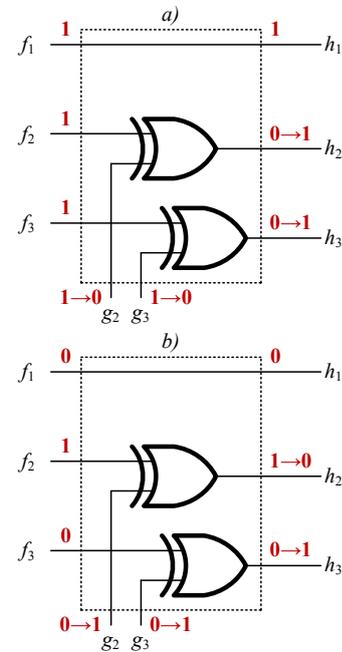


Fig. 8. Examples of error on outputs of Boolean Complement Segment: a) detectable; b) non-detectable.

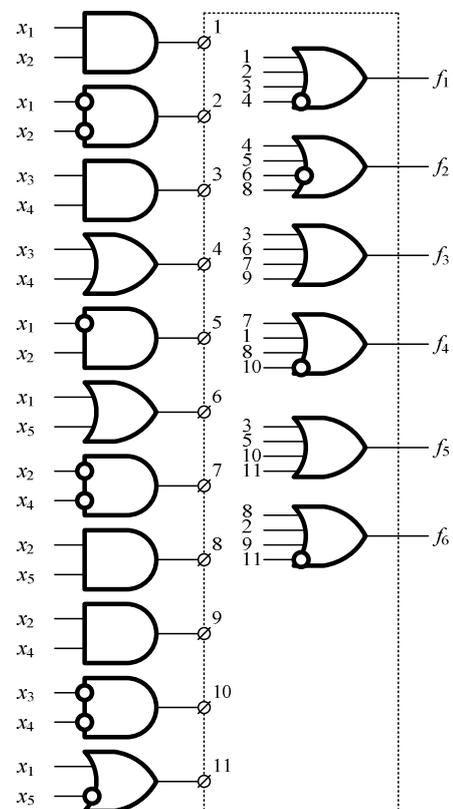


Fig. 9. Combinational circuit with six outputs.

IV. CONCLUSION

Boolean Complement Method has flexible tools for the cause of logical complement option which helps to synthesize those CED systems with reduced structural abundance compared to Check Bits Calculation Method and conventional Back-Up Structure. Nevertheless, as is shown in present article via implementation of additional segment per main with control logical segments into the CED system, several limits should be added for the status quo. Otherwise, the whole diagnostic system should not obtain the self-checking function.

Presented statements of this article were completed for the simple event of 1/3-code CED system relevance as a basic kit. More then, that formulated statements should be easy to generalize in the event of any r/n -codes application.

REFERENCES

- [1] M. Goessel, and S. Graf "Error Detection Circuits", London: McGraw-Hill, 1994, 261 p.
- [2] D.K. Pradhan "Fault-Tolerant Computer System Design", New York: Prentice Hall, 1996, 560 p.
- [3] R. Dobias, and H. Kubatova "FPGA Based Design of the Railway's Interlocking Equipments", Euromicro Symposium on Digital System Design, 2004 (DSD 2004), 31 August – 3 September 2004, Rennes, France, DOI: 10.1109/DSD.2004.1333312.
- [4] P.K. Lala "Principles of Modern Digital Design", New Jersey: John Wiley & Sons, 2007, 419 p.
- [5] R. Ubar, J. Raik, and H.-T. Vierhaus "Design and Test Technology for Dependable Systems-on-Chip (Premier Reference Source)", Information Science Reference, Hershey – New York, IGI Global, 2011, 578 p.
- [6] M. Goessel, V.I. Sapozhnikov, V. Sapozhnikov, and A. Dmitriev "A New Method for Concurrent Checking by Use of a 1-out-of-4 Code", Proceedings of the 6th IEEE International On-line Testing Workshop, 3-5 July 2000, Palma de Mallorca, Spain, pp. 147-152.
- [7] S. Mitra, and E.J. McCluskey "Which Concurrent Error Detection Scheme to Choose?", Proceedings of International Test Conference, 2000, USA, Atlantic City, NJ, 03-05 October 2000, pp. 985-994, doi: 10.1109/TEST.2000.894311.
- [8] V.V. Sapozhnikov, V.I. Sapozhnikov, A. Morozov, G. Osadchi, and M. Gossel "Design of Totally Self-Checking Combinational Circuits by Use of Complementary Circuits", Proceedings of East-West Design & Test Workshop, Yalta, Ukraine, 2004, pp. 83-87.
- [9] L.-T. Wang, C.-W. Wu, and X. Wen "VLSI Test Principles and Architectures: Design for Testability", USA, San Francisco, Morgan Kaufmann Publishers, 2006, 777 p.
- [10] J. Borecký, M. Kohlík, and H. Kubátová "Parity Driven Reconfigurable Duplex System", Microprocessors and Microsystems, 2017, Vol. 52, pp. 251-260, doi: 10.1016/j.micpro.2017.06.015.
- [11] M. Goessel, A.V. Morozov, V.V. Sapozhnikov, and V.I. Sapozhnikov "Checking Combinational Circuits by the Method of Logic Complement", Automation & Remote Control, 2005, vol. 66, issue 8, pp. 1336-1346.
- [12] M. Nicolaidis, and Y. Zorian "On-Line Testing for VLSI – A Compendium of Approaches", Journal of Electronic Testing: Theory and Application, 1998, vol. 12, issue 1-2, pp. 7-20, doi: 10.1023/A:1008244815697.
- [13] V.V. Sapozhnikov, A. Morozov, V.I. Sapozhnikov, and M. Goessel "Concurrent Checking by Use of Complementary Circuits for «1-out-of-3» Codes", 5th International Workshop IEEE DDECS 2002, Brno, Czech Republic, April 17-19, 2002.
- [14] M. Goessel, A.V. Morozov, V.V. Sapozhnikov, and V.I. Sapozhnikov "Logic Complement, a New Method of Checking the Combinational Circuits", Automation and Remote Control, 2003, vol. 1, issue 1, pp. 153-161.
- [15] M. Goessel, V. Ocheretny, E. Sogomonyan, and D. Marienfeld "New Methods of Concurrent Checking: Edition 1", Dordrecht: Springer Science+Business Media B.V., 2008, 184 p.
- [16] S.K. Sen "A Self-Checking Circuit for Concurrent Checking by 1-out-of-4 code with Design Optimization using Constraint Don't Cares", National Conference on Emerging trends and advances in Electrical Engineering and Renewable Energy (NCEEERE 2010), Sikkim Manipal Institute of Technology, Sikkim, held during 22-24 December, 2010.
- [17] D. Efanov, V. Sapozhnikov, and V.I. Sapozhnikov "Method of Self-Checking Concurrent Error Detection System Development Based on Constant-Weight Code «2-out-of-4»", Proceedings of 3ed International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), St. Petersburg, Russia, May 16-19, 2017, doi: 10.1109/ICIEAM.2017.8076374.
- [18] V. Sapozhnikov, V.I. Sapozhnikov, D. Efanov, A. Bliudov, and D. Pivovarov "Combinational Circuit Check by Boolean Complement Method Based on «1-out-of-5» Code", Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 89-94, doi: 10.1109/EWDTS.2017.8110076.
- [19] J.L.A. Huches, E.J. McCluskey, and D.J. Lu "Design of Totally Self-Checking Comparators with an Arbitrary Number of Inputs", IEEE Transactions on Computers, 1984, vol. C-33, no. 6, pp. 546-550.
- [20] G.P. Aksyonova "Necessary and Sufficient Conditions for the Design of Totally Checking Circuits of Compression by Modulo 2", Automation & Remote Control, 1979, vol. 40, issue 9, pp. 1362-1369.
- [21] D.K. Das, S.S. Roy, A. Dmitriev, A. Morozov, and M. Gossel "Constraint Don't Cares for Optimizing Designs for Concurrent Checking by 1-out-of-3 Codes", Proceedings of the 10th International Workshops on Boolean Problems, Freiberg, Germany, September, 2012, pp. 33-40.
- [22] D. Efanov, V. Sapozhnikov, and V.I. Sapozhnikov "Methods of Organization of Totally Self-Checking Concurrent Error Detection System on the Basis of Constant-Weight «1-out-of-3»-Code", Proceedings of 14th IEEE East-West Design & Test Symposium (EWDTS'2016), Yerevan, Armenia, October 14-17, 2016, pp. 117-125, doi: 10.1109/EWDTS.2016.7807622.
- [23] V.V. Sapozhnikov, V.I. Sapozhnikov, D.V. Efanov, and D.V. Pivovarov "Synthesis of Concurrent Error Detection Systems of Multioutput Combinational Circuits Based on Boolean Complement Method" (in Russian), Tomsk State University Journal of Control and Computer Science, 2017, issue 4, pp. 69-80, doi: 10.17223/19988605/41/9.
- [24] E.S. Sogomonyan, and M. Gossel "Design of Self-Testing and On-Line Fault Detection Combinational Circuits with Weakly Independent Outputs", Journal of Electronic Testing: Theory and Applications, 1993, vol. 4, issue 4, pp. 267-281, doi:10.1007/BF00971975.
- [25] A. Morosow, V.V. Sapozhnikov, V.I. Sapozhnikov, and M. Goessel "Self-Checking Combinational Circuits with Unidirectionally Independent Outputs", VLSI Design, 1998, vol. 5, issue 4, pp. 333-345, doi: 10.1155/1998/20389.
- [26] D. Efanov, V. Sapozhnikov, and V.I. Sapozhnikov "Synthesis of Self-Checking Combinational Devices Based on Allocating Special Groups of Outputs", Automation and Remote Control, 2018, issue 9, pp. 1607-1618, doi: 10.1134/S0005117918090060.

Use of Natural Information Redundancy in On-Line Testing of Computer Systems and their Components

Oleksandr Drozd
Department of Computer Intelligent
Systems and Networks
Odessa National Polytechnic
University
Odessa, Ukraine
drozd@ukr.net

Anatoliy Sachenko^{1,2}
¹Kazimierz Pulaski University of
Technology and Humanities in Radom,
Radom 26-600, Poland,
sachenkoa@yahoo.com
²Ternopil National Economic University
Ternopil, 46027, Ukraine

Svetlana Antoshchuk
Department of Computer Intelligent
Systems and Networks
Odessa National Polytechnic
University
Odessa, Ukraine
asgonpu@gmail.com

Julia Drozd
Department of Computer Systems
Odessa National Polytechnic University
Odessa, Ukraine
dea_lucis@ukr.net

Mykola Kuznietsov
Department of Computer Systems
Odessa National Polytechnic University
Odessa, Ukraine
koliaodessa@mail.ru

Abstract—This paper is devoted to the problem of on-line testing the digital circuits in the use of natural information redundancy for results calculated during the performing arithmetic operations on approximate data. Those data are represented usually in floating-point formats. It's shown the dominant development of computer systems in critical applications that support approximate calculations performed on the measurement results obtained from sensors. Information redundancy is considered as the basis for solving analysis issues, including the on-line testing of digital circuits. The existence of natural information redundancy in the result codes of all arithmetic operations on approximate data is proved. These codes contain forbidden values that are calculated under the action of faults. Another type of natural information redundancy inherent in data formats in unused positions is considered as well. The role of natural information redundancy in a version form for increasing the checkability of the circuits and trustworthiness of result in the FPGA components of safety-related systems is noted.

Keywords—*Natural Information Redundancy, On-Line Testing, Approximate Data Processing, Fixed and Floating-Point Format, Resource Approach, Safety-Related System, Version Information Redundancy, FPGA Design*

I. INTRODUCTION

On-line testing occupies an important place in ensuring the functionality of computer systems and their digital components [1, 2]. Methods and means for on-line testing are aimed at checking the trustworthiness of the results calculated at the output of digital circuits. This goal was originally formulated as the detection of digital circuit faults in the operating mode, since the methods and means for on-line testing began to develop within the framework of the exact data model, which was reflected in the theory and practice of constructing totally self-checking schemes [3, 4].

For exact data, i.e. integers in nature (the numbers of the elements of the sets), these goals of the on-line testing do not differ. Indeed, an error detected in a calculated result simultaneously indicates both the presence of a fault and an incorrect result, which for exact data consists entirely of most significant bits and is therefore unreliable.

Computer systems demonstrate the dominant development in the processing of approximate data, when it is impossible to further ignore the structure of the

approximate number consisting of the most and least significant bits. Faults in them cause errors that are respectively essential and inessential errors for the trustworthiness of the result [5, 6].

Approximate data are measurements and, as a rule, are presented and processed in floating-point formats. In the first personal computers, such processing was carried out in the optional Intel 287/387 coprocessors. The next Pentium family provides accelerated pipelined processing of approximate data, and a modern graphics processor contains thousands of floating-point pipelines used for performance of parallel calculations on CUDA technology [7, 8].

From the position of the resource approach, this development of personal computers reflects the general tendency of structuring the resources, i.e. models, methods and means, under the peculiarities of the natural world, among which its parallelism and fuzziness turned out to be most apparent in computing [9, 10].

The resource approach notes three levels of development of the resources in the process of their integration into the natural world: replication, diversification and self-sufficiency [11].

Replication, as the simplest form of production and parallelization, will always be selected with open resource niches – ecological, technological, market, etc. At the level of replication, integration into the natural world is carried out at the expense of productivity. In the natural world, it occurs under the slogan: “Give birth more than die”. With the closure of resource niches, clones can only survive by showing features, that is, becoming individuals, rising to a level of diversification, where survival is at the expense of authenticity, trustworthiness, i.e. adequacy to the natural world. At this level, we solve problems of analysis, including diagnostic problems that for the decision require the use of information redundancy, which reflects the fuzziness of the natural world [12, 13].

We observe filling of the natural world with objects of the increased risk – powerful power plants, power grids, high-speed transport, various types of weapons. This process changes the benchmarks in the development of computer systems from productivity to trustworthiness, defining them as instrumentation and control safety-related systems, aimed

at ensuring functional safety of both the system and the control object to prevent accidents and reduce their consequences [14, 15].

In these systems, the operating mode is diversified, dividing into normal and emergency. The goal of on-line testing, which plays a key role in the operational assessment of the state of the system, is also diversified. In emergency mode, results are monitored with their check for trustworthiness. Normal mode is used as a test for clearing digital circuits from faults, where elements of testability and self-testing are important [16, 17]. The circuits have to show faults in the normal mode, i.e. to be checkable, and to mask faults in emergency mode for obtaining reliable results.

It should be noted that safety-related systems, like cyber-physical systems, and Internet of Things systems, receive initial data from sensors, i.e. get measurement results that relate to approximate data [18, 19]. Thus, the position of the on-line testing is enhanced in the processing of approximate data for new directions in development of computer systems and information technology.

Self-sufficiency determines the level of development goal. Resources strive in their development for self-sufficiency. In particular, our models, in which methods and means are being developed, rise to an understanding of the natural resources that are already embedded in the models, methods and means like the basis for the development of their self-sufficiency.

In confirmation to it, the results of arithmetic operations demonstrate self-sufficiency in the presence of natural information redundancy, sufficient for the implementation of on-line testing with the use of the forbidden values in the result code.

The paper is devoted to the analysis of natural information redundancy as an inherent property of the results of arithmetic operations with approximate data. There are some examples of the use of natural information redundancy for the on-line testing of digital circuits. This paper reveals the universal nature of this resource, its availability in all applications for on-line testing in the processing of approximate data. Section 2 proves the existence of natural information redundancy in the results of all arithmetic operations with approximate data. Here information redundancy is considered in the traditional form of the presence of forbidden values in the result code. Section 3 draws attention to the natural information redundancy generated by fixed and floating-point data formats. Section 4 shows the natural information redundancy in version form for program code of FPGA projects in critical applications.

II. UNIVERSAL NATURE OF NATURAL INFORMATION REDUNDANCY IN THE APPROXIMATE RESULTS

Traditionally, information redundancy of a result code is interpreted as the presence of forbidden values in it, which cannot be calculated in a properly operating digital circuit.

To obtain information redundancy, the binary number, as a rule, is supplemented with a check code, which lengthens this number. For example, in residue checking, which is widely used for on-line testing of arithmetic operations, the check code by modulo three complements the original n -bit number with two bits to the $(n + 2)$ -bit number [20, 21].

The values of the original number with the corresponding check code form a set of allowed words that can be calculated with the correct functioning of the scheme. This set contains 2^n words.

However, an $(n + 2)$ -bit number can take 2^{n+2} different values, of which only 2^n words are allowed. The remaining words make up a set of forbidden words. For $n = 8$, the checking by modulo three generates 256 allowed and $1024 - 256 = 768$ forbidden words.

On-line testing is based on the distinction between allowed and forbidden words. Identification of a forbidden word is unambiguously associated with an error caused by a circuit fault, and determines the result to be unreliable.

Natural information redundancy is the presence of forbidden words in the result code, considered without the check code.

Specific examples of results, the codes of which contain forbidden words, indicate the existence of natural information redundancy. However, the presence of forbidden words in the results of processing approximate data is not a special case, but a general rule.

The universal nature of the natural information redundancy in the results of arithmetic operations on approximate data follows from the following two statements:

- 1) The code of a product has natural information redundancy.
- 2) Multiplication is a key operation of approximate calculations.

The first statement is based on the same length of the input and output words in the multiplication operation and on the commutative law for it. The input word consists of codes of factors. For n -bit binary factors, the input word has a size of $2n$ bits. The output word, i.e. the complete product code also has a length of $2n$ bits. Therefore, the input word and the output word take the same number of 2^{2n} values. However, the commutative law states that the product does not change from changing the places of the factors. Thus, the same product corresponds to different input words with interchanged factors. Therefore, any output words have no match in the set of input words and therefore belong to the forbidden.

For example, for $n = 2$, the factors take the values 0, 1, 2, 3, and their products, 0, 1, 2, 3, 4, 6, 9, form a set of seven allowed values. The product code has 4 bits in size and takes values from 0 to 15, of which the numbers 5, 7, 8, 10, 11, 12, 13, 14 and 15 form a set of forbidden values.

The second statement follows from the use of the normal form in the representation of numbers when they are written in floating-point formats. This representation contains the mantissa, the exponent, by default implies the base of the number system and combines these elements using the multiplication operation [22, 23].

The presence of a multiplication operation in the floating-point number itself determines the use of multiplication, in one form or another, in all operations with mantissas. The results of these operations inherit the properties of the product, i.e. possess natural information redundancy.

The residue checking does not distinguish between allowed and forbidden values of the product code and the code of any other result of the processing of approximate data. He relates all the values of a product to allowed words and spends resources on creating information redundancy, ignoring its natural form. Diversification of words in the code of an approximate result by dividing them into allowed and forbidden allows you to perform on-line testing at the expense of internal resources of arithmetic operations with approximate data. Successful examples of such use are known for multiplication of binary codes and their squaring [24].

In case of multiplication of two mantissas, the binary code of the product contains the forbidden values of a type $P = kC$, where $k = 2^{n-1}, \dots, 2^n - 1$, $C = 2^n + 1$ – a prime number, for example, 17, 257 and 65537 for $n = 4, 8$ and 16, respectively. Really, number C and multiple to it numbers cannot be the product of two n -bit binary numbers as they have no decomposition on two factors with the size smaller than $n + 1$. These forbidden values easily are being identified as they form the code with repetition, when a younger half of the product coincides with the senior half. The Z error is being detected in case of performing a condition of $A + B + Z = P$, where A and B are binary codes of mantissas of the normalized factors: $A = 2^{n-1}, \dots, 2^n - 1$, $B = 2^{n-1}, \dots, 2^n - 1$.

The errors typical for the iterative array multiplier distort the product on the weight of any one its bit $U = n, \dots, 2n - 1$ and they can be described as $Z = \pm 2^U$. We experimentally showed that all such errors can be detected for $n = 4, 8$ and 16. In the course of the experiments, the specially developed program determined A, B and k values for which the condition of error detection is satisfied for each U . For example, for $n = 4$, the error $Z = 2^5$ or $Z = -2^5$ in bit $U = 5$ is detected on the condition of $8 \times 13 + 32 = 8 \times 17$ or $12 \times 14 - 32 = 8 \times 17$ in case of $A = 8, B = 13, k = 8$ or $A = 12, B = 14, k = 8$, respectively.

The S result of squaring shows natural information redundancy in values of the residue by the modulo. For example, $S \bmod 3 \neq 2$. The error is detected when performing a condition $(S + 1) \bmod 3 = 0$.

We experimentally showed that all Z errors typical for the matrix scheme of a squarer with word size of $n = 8, 16$ and 32 can be detected. The error detection scheme defines the residue $1 = 01_2$, or $2 = 10_2$ during the correct work and $+0 = 00_2$ or $-0 = 11_2$ in case of error detection. The scheme belongs to totally self-checking.

III. NATURAL INFORMATION REDUNDANCY OF DATA FORMATS

Statements about the only form of information redundancy based on the existence of forbidden values are known [25].

At the same time, data formats demonstrate a different form of information redundancy, which is characterized as natural and has properties important for on-line testing of the digital circuits in the processing of approximate data.

Modern fixed and floating-point formats can be opposed to each other by the positions at which codes of numbers are placed. In the case of a fixed point, the low bits of the numbers occupy the lower positions of the format. On the contrary, the floating-point format is characterized by the

placement of the most significant bits of the number in higher positions.

The size of the fixed-point format is selected based on the representation of the longest number. When writing numbers of shorter length, the highest format positions are not used. Such natural information redundancy is the basis of logarithmic checking, which could not compete with a method of the residue checking within the framework of the exact data model [26]. At the same time, this method demonstrates an important property of more probable error detection in higher positions than in younger ones. Such a distinction between essential and inessential errors based on the natural information redundancy of data formats is particularly important for the on-line testing of approximate calculations [27].

In floating-point formats, the unit value of the most significant bit of the mantissa is fixed. The following zero values form the natural information redundancy of floating-point formats.

We developed the program in Delphi 10 Seattle demo-version [28] for a pilot study of logarithmic checking on the example of the fixed-point multiplier. The Ac check code of number A is defined by the number of bits of its significant part and $Ac = 0$ in case of $A = 0$ [29]. Then for $A, B > 0$, the product $P = A \times B$ can be checked as $Pc = Pc^*$ or $Pc = Pc^* - 1$, where $Pc^* = Ac + Bc$. In case of $A, B \geq 0$, $Pc^* = Ac Bz + Bc Az$, where $Az = 0$ for $A = 0$ and $Az = 1$ for $A > 0$; $Bz = 0$ for $B = 0$ and $Bz = 1$ for $B > 0$. It follows from the following inequalities: $2^{Ac-1} \leq A < 2^{Ac}$, $2^{Bc-1} \leq B < 2^{Bc}$, $2^{Pc-1} \leq A < 2^{Pc}$ and definitions of the lower (top) bound of the product as a product of the lower (top) bounds of factors.

The iterative array multiplier contains $n \times (n - 1)$ matrix of operational elements. Multiplication is carried out for one clock cycle on the factors generated in a random way. In each clock cycle, fault of short circuit between two any points of the scheme in accidentally chosen operational element is injected. In an experiment, the word size $n = 8, \dots, 15$ and the number $n_M \leq n$ of most significant bits is determined.

The program panel with results of an experiment for $n = 8$ and $n_M = n$ is shown in Fig. 1.

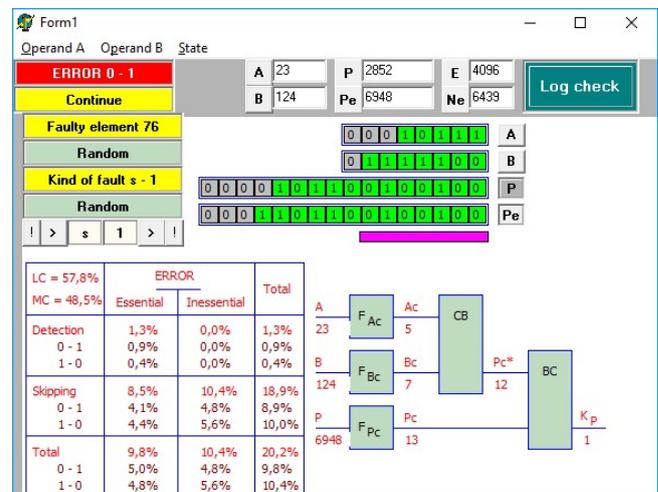


Fig. 1. Results of experiment

The multiplier has calculated both incorrect result $Pe = 6948$, $Pe = 0001011111110000_2$ and its check code $Pc = 13$ which is compared to the Pc^* code. The error $E = 4096$ was caused by fault of short circuit between the s sum and the level of logical unit in the full adder of operational element 76 (line 7 and column 6 of multiplier matrix). This error is detected when checking $Pc > Pc^*$. The experiment also shows probabilities of detection and the skipping of essential and inessential, positive (0 - 1) and negative (1 - 0) errors. Besides, the trustworthiness of logarithmic (LC) and residue (MC) checking is estimated.

The trustworthiness is calculated by the formula $T = P_{ED} + P_{ID}$, where $P_{ED} = P_E P_D$ and $P_{ID} = (1 - P_E)(1 - P_D)$ are probability of essential error detection and the probability of inessential error skipping; P_E and P_D – the probability of an essential error and probability of error detection, respectively. Residue checking showed detection of all errors, i.e. $P_D = 1$, and its trustworthiness is defined as $T_{MC} = P_E$ [24].

The experiment showed the following values of probabilities: $P_{ED} = 1.3 / 20.2 = 6.4\%$, $P_{ID} = 10.4 / 20.2 = 51.5\%$, $P_E = 9.8 / 20.2 = 48.5\%$. Trustworthiness of logarithmic and residue checking is estimated as 57.8% and 48.5%, respectively.

IV. VERSION FORM OF THE NATURAL INFORMATION REDUNDANCY

In systems of critical application, functional safety is ensured through the use of fault-tolerant solutions [30], among which multi-version technologies are of particular importance [31]. They allow to resist to common cause failures and accumulation of the hidden faults [32, 33], to raise integrity [34] and also at the same time a checkability of FPGA projects and trustworthiness of results in normal and emergency mode, respectively [35, 36].

FPGA designing with the LUT-oriented architecture widely is used by in critical applications [37, 38], creating conditions for development of natural information redundancy in a version form when for the ready FPGA project, the set of versions of the program code with various properties is generated.

The LUT unit generates function from n arguments. For $n = 4$, the LUT unit has 4 address inputs of A, B, C, D on which a, b, c, d arguments forming the $dcb a_2$ address code, 16 bits of memory and one output [39, 40]. The sequence of LUT codes forms a program code of the FPGA project. The version redundancy of this code is based on an opportunity to create its versions for the same hardware implementation of the project. Versions are generated at the level of ordered pair LUT1 - LUT2 of the LUT units connected among themselves: LUT1 output bit transfers to the address input of the LUT2 unit.

This bit can be transferred by a direct or inverse value, forming two versions of a program code. The inverse value of bit forms by inversion of bits of the LUT1 memory. This inversion on an address input of LUT2 unit is compensated by change of places of its bits of memory.

The contrast of the described versions can be effectively used for opposition to faults of short circuit of the next address inputs of LUT [41]. Really, inversion of one of them interchanges the position of a set of input data on which this fault is shown in the form of an error or is masked. It allows to find versions which promote fault detection on some

input data and mask on others, i.e. to raise a checkability of circuits and trustworthiness of results in the normal and emergency modes of safety-related systems.

The example of versions with such opportunities in case of a direct and inverse address input D is shown in Fig. 2.

dc	Version 1				Version 2							
	11	10	01	00	11	10	01	00				
00	0	1	1	0	0	1	1	0	1	0	0	0
01	0	0	0	1	0	1	1	0	0	1	0	0
10	1	0	0	0	0	1	1	0	0	1	0	0
11	0	1	0	0	0	1	0	0	0	0	0	1

Correct circuit The C-D fault Correct circuit The C-D fault

■ – the bits addressed in a normal mode
■ – the bits addressed in an emergency mode

Fig. 2. LUT2 unit in two versions with direct and inverse D address input

Each version shows LUT2 unit with the correct circuit and in case of short circuit between address inputs C and D. The fault copies the line $dc = 00_2$ into lines $dc = 01_2$ and $dc = 10_2$. Inversion on the D address input is compensated by change of places of the lines $dc = 0X_2$ and $dc = 1X_2$ where $X = 0$ or $X = 1$. This example is reviewed for a case when bits with addresses $XX0X_2$ and $XX1X_2$ are addressed in normal and emergency mode, respectively.

The checkability of the circuit and trustworthiness of result in mode U is estimated by the formulas $C_U = Z_U / Y_U$ and $T_U = 1 - C_U$, where $U = N$ and $U = E$ for normal and emergency mode, respectively, Z_U and Y_U – the number of the distorted bits and their total in mode U .

An example shows the following results: $Z_N = 3$, $Y_N = 8$, $C_N = 37.5\%$, $T_N = 62.5\%$, $Z_E = 3$, $Y_E = 8$, $C_E = 37.5\%$, $T_E = 62.5\%$ and $Z_N = 4$, $Y_N = 8$, $C_N = 50\%$, $T_N = 50\%$, $Z_E = 1$, $Y_E = 8$, $C_E = 12.5\%$, $T_E = 87.5\%$ for versions 1 and 2, respectively. Thus, transition from version 1 to version 2 raises at the same time both a checkability of the circuit and trustworthiness of result respectively in normal and emergency mode. The checkability improves for 12.5%, and reliability – for 25%.

V. CONCLUSIONS

Information redundancy is the basis for solving the analysis issues, including the issues of on-line testing the digital circuits in computer systems and their components. Information redundancy is created in the form of forbidden values by forming the check codes of numbers, which requires the certain hardware and time resources. Existing examples of using the natural information redundancy are not regular. The orientation of modern computer systems on critical applications and processing the approximate data increases the importance of on-line testing the digital circuits in relation to checking the approximate calculations. In these circumstances, a multiplication becomes the key operation, and the natural information redundancy inherent in the product gets the universal character, extending to the results of all arithmetic operations on mantissas. Thus, the on-line testing obtains a natural resource for the dominant processing of approximate data.

Natural information redundancy also takes a place in fixed and floating-point formats in the form of unused positions. Logarithmic checking, using this form of redundancy, demonstrates an important feature of more likely detection of essential errors compared to inessential ones.

The version form of natural information redundancy allows to reach at the same time a high checkability of FPGA projects and trustworthiness of results respectively in normal and emergency mode of safety-related systems.

REFERENCES

- [1] M. Nicolaidis, and Y. Zorian, "On-Line Testing for VLSI – a Compendium of Approaches. Electronic Testing: Theory and Application." JETTA, 1998, vol. 12, pp. 7-20.
- [2] A. Drozd, M. Lobachev, J. Drozd, "The problem of on-line testing methods in approximate data processing," 12th IEEE International On-Line Testing Symposium, Como, Italy, 2006, pp. 251-256. DOI:10.1109/IOLTS.2006.61
- [3] C. Metra, L. Schiano, M. Favalli, B. Ricco, "SelfChecking scheme for the on-line testing of power supply noise," Design, Automation and Test in Europe Conference, Paris, France, 2002, pp. 832-836.
- [4] D. Efanov, V. Sapozhnikov, VI. Sapozhnikov, "Applications of modular summation codes to concurrent error detection systems for combinational boolean circuits," Automation and Remote Control, 2015, vol. 76, issue 10, pp. 1834-1848.
- [5] H. Kekre, D. Mishra, R. Khanna, S. Khanna, A. Hussaini, "Comparison between the basic LSB Replacement Technique & Increased Capacity of Information Hiding in LSB's Method for Images," International Journal of Computer Applications, 2012, vol. 45, no. 1, pp. 33-38.
- [6] S. O. Akinola, A. A. Olatidoye, "On the image quality and encoding times of LSB, MSB and combined LSB-MSB steganography algorithms using digital images," International Journal of Computer Science & Information Technology, 2015, vol. 7, no 4, pp. 79-91.
- [7] R. Hiromoto, "Parallelism and complexity of a small-world network model," International Journal of Computing, 2016, vol. 15, issue 2, pp. 72-83.
- [8] NVIDIA CUDA Compute Unified Device Architecture. Programming Guide / Version 1.0, NVIDIA Corporation, 2007.
- [9] V. Opanasenko, S. Kryvyi, "Synthesis of multilevel structures with multiple outputs," CEUR Workshop Proceeding, 2016, vol. 1631, pp. 32-37.
- [10] Y. P. Kondratenko, L. P. Klymenko, E. Y. M. Al Zu'bi, "Structural Optimization of Fuzzy Systems' Rules Base and Aggregation Models," Kybernetes, 2013, vol. 42, no. 5, pp. 831-843.
- [11] J. Drozd, A. Drozd, S. Antoshchuk, V. Kharchenko, "Natural Development of the Resources in Design and Testing of the Computer Systems and their Components," IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Berlin, Germany, 2013, pp. 233-237. DOI: 10.1109/IDAACS.2013.6662656
- [12] Synopsys. DWFC Flexible Floating Point Overview, no. August, pp. 1-6, 2016.
- [13] A. Drozd, J. Drozd, S. Antoshchuk, V. Antonyuk, K. Zashcholkin, M. Drozd, O. Titomir, "Green Experiments with FPGA. In book: Green IT Engineering: Components, Networks and Systems Implementation, V. Kharchenko, Y. Kondratenko, J. Kacprzyk (Eds.), vol. 105. Berlin, Heidelberg: Springer International Publishing, 2017, pp. 219-239. DOI: 10.1007/978-3-319-55595-9_11
- [14] V. Kharchenko, A. Gorbenko, V. Sklyar, C. Phillips, "Green Computing and Communications in Critical Application Domains: Challenges and Solutions," IX International Conference of Digital Technologies, Zhilina, Slovak Republic, 2013, pp. 191-197.
- [15] IEC 61508-1:2010. Functional Safety of Electrical / Electronic / Programmable Electronic Safety Related Systems – Part 1: General requirements. Geneva: International Electrotechnical Commission, 2010.
- [16] A. Matrosova, E. Nikolaeva, D. Kudin, V. Singh, "PDF testability of the circuits derived by special covering ROBDDs with gates," IEEE East-West Design and Test Symposium, EWDTs, 2013.
- [17] V. A. Romankevich, "Self-testing of multiprocessor systems with regular diagnostic connections," Automation and Remote Control, 2017, vol. 78, issue 2, pp. 289-299.
- [18] V. Hahanov, E. Litvinova, S. Chumachenko, "Cyber Physical Computing for IoT-driven Services," Springer, 2017.
- [19] D. Maevsky, A. Bojko, E. Maevskaya, O. Vinakov, L. Shapa, "Internet of things: Hierarchy of smart systems," 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017, pp. 821-827.
- [20] V. Hema, M. G. Durga, "Data Integrity Checking Based on Residue Number System and Chinese Remainder Theorem in Cloud," International Journal of Innovative Research in Science, Engineering and Technology, 2014, vol. 3, special issue 3.
- [21] A. V. Drozd, M. V. Lobachev, W. Hassonah, "Hardware check of Arithmetic Devices with Abridged Execution of Operations," European Design & Test Conference (ED & TC 96), Paris, France, 1996, p. 611. DOI: 10.1109/EDTC.1996.494375
- [22] ANSI/IEEE Std 754-1985. IEEE Standard for Binary Floating-Point Arithmetic, 1985.
- [23] IEEE Std 754™-2008 (Revision of IEEE Std 754-1985) IEEE Standard for Floating-Point Arithmetic. IEEE 3 Park Avenue New York, NY 10016-5997, USA, 2008.
- [24] A. Drozd, S. Antoshchuk, "New on-line testing methods for approximate data processing in the computing circuits," 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Prague, Czech Republic, 2011, pp. 291-294. DOI: 10.1109/IDAACS.2011.6072759
- [25] A. K. Sushkevich, "Theory of numbers," Kharkiv, Kharkiv State University, 1956.
- [26] F. Sellers, M. Hsiao, L. Beamson, "Error Detecting Logic for Digital Computers," New-York: Mc GRAW-HILL, 1968.
- [27] A. Drozd, M. Lobachev, "Efficient On-line Testing Method for Floating-Point Adder," IEEE Design, Aut. and Test in Europe (DATE), 2001, pp. 307-311. DOI: 10.1109/DATE.2001.915042
- [28] Delphi 10 Seattle: Embarcadero, 2015. <https://www.embarcadero.com/ru/products/delphi>
- [29] A. Drozd, R. Al-Azzeh, J. Drozd, M. Lobachev, "The logarithmic checking method for on-line testing of computing circuits for processing of the approximated data," Euromicro Symposium on Digital System Design, Rennes, France, 2004, pp. 416-423.
- [30] I. Atamanyuk, Y. Kondratenko, "Computer's analysis method and reliability assessment of fault-tolerance operation of information systems," CEUR Workshop Proceedings, 2015, vol. 1356, pp. 507-522.
- [31] H. Asad, I. Gashi, "Diversity in Open Source Intrusion Detection Systems," Computer Safety, Reliability, and Security Lecture Notes in Computer Science, 8666, Springer, 2014, pp. 267-281.
- [32] A. Drozd, M. Drozd, V. Antonyuk, "Features of Hidden Fault Detection in Pipeline Components of Safety-Related System," CEUR Workshop Proceedings, 2015, vol. 1356, pp. 476-485.
- [33] IEC 62340:2007. Nuclear power plants – Instrumentation and control systems important to safety – Requirements for coping with common cause failure. Geneva: International Electrotechnical Commission, 2007.
- [34] K. Zashcholkin, O. Ivanova, "The control technology of integrity and legitimacy of LUT-oriented information object usage by self-recovering digital watermark," CEUR Workshop Proceedings, 2015, vol. 1356, pp. 498-506.
- [35] A. Drozd, M. Drozd, M. Kuznietsov, "Use of Natural LUT Redundancy to Improve Trustworthiness of FPGA Design," CEUR Workshop Proceedings, 2016, vol. 1614, pp. 322-331.
- [36] A. Drozd, M. Drozd, O. Martynyuk, M. Kuznietsov, "Improving of a Circuit Checkability and Trustworthiness of Data Processing Results in LUT-based FPGA Components of Safety-Related Systems," CEUR Workshop Proceedings, 2017, vol. 1844, pp. 654-661.
- [37] S. F. Tyurin, A. V. Grekov, O.A. Gromov, "The principle of recovery logic FPGA for critical applications by adapting to failures of logic elements," World Applied Sciences Journal, 2013, pp. 328-332.
- [38] K. Zashcholkin, O. Ivanova, "LUT-object integrity monitoring methods based on low impact embedding of digital watermark," 14th International Conference "Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering" (TCSET-2018), 2018, pp. 519-523.
- [39] Intel Cyclone FPGA series, <https://www.intel.com/content/www/us/en/products/programmable/cyclone-series.html>, last accessed 2019/04/26.
- [40] Intel FPGA Architecture, <https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/wp-01003.pdf>, last accessed 2019/04/26.
- [41] W. Pleskacz, M. Jenihhin, J. Raik, M. Rakowski, R. Ubar, W. Kuzmicz, "Hierarchical Analysis of Short Defects between Metal Lines in CMOS IC," 11th Euromicro Conference on Digital System Design Architectures, Methods and Tools, Parma, Italy, 2008, pp. 729-734.

Self-Dual Complement Method up to Constant-Weight Codes for Arrangement of Combinational Logical Circuits Concurrent Error-Detection Systems

Dmitry Efanov,
DSc, Professor at “Automation, Remote
Control and Communication
on Railway Transport”,
Russian University of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

Valery Sapozhnikov,
DSc, professor at “Automation and Remote
Control on Railways” Department,
Emperor Alexander I St. Petersburg State
Transport University,
St. Petersburg, Russia
port.at.pgups@gmail.com

Vladimir Sapozhnikov,
DSc, professor at “Automation and Remote
Control on Railways” Department,
Emperor Alexander I St. Petersburg State
Transport University,
St. Petersburg, Russia
at.pgups@gmail.com

German Osadchy,
Technical Director of Scientific and Technical Center “Integrated
Monitoring Systems” LLC,
St. Petersburg, Russia
osgerman@mail.ru

Dmitry Pivovarov,
PhD student at “Automation and Remote Control on Railways”
Department, Emperor Alexander I
St. Petersburg State Transport University,
St. Petersburg, Russia
pivovarov.d.v.spb@gmail.com

Abstract—Authors of this article are being recommended a new approach to concurrent checking method arrangement, composed of main features of Self-Dual Complement Method together with constant-weight code. According to analysis we may state that in the event of single control application based on self-dual function characteristic either via being arranged digital word to constant-weight code, some errors of comprised circuit may not be detected. Unification procedure while functional control of two characteristics helps us to advance the whole concurrent error-detection (CED) system regarding errors identification. In this article we are presenting realization method of concurrent checking in accordance with Self-Dual Complement Method up to constant-weight code. It was demonstrated that to ensure supervision characteristic per two symptoms of the system we may apply constant-weight code ‘r out of 2r’ (r – weight value of the digital word of the constant-weight code). In this case, the developed via authors method of CED system arrangement considered self-dual features for each function of the constant-weight code words. The most effective instrument in this approach is reckoned code ‘2 out of 4’ with simple checker structure for the purpose of the entire check out of few coded combinations. In the event of multiple outputs circuits those exit ports should be divided per four groups for each of those matrices and outputs of separate control circuits must be unified upon self-checkable comparator output. The abovementioned option allowed us to widen the aforementioned way to synthesis of combinational circuits under supervision.

Keywords—self-dual complement, coded word, constant-weight code, concurrent error-detection system, Boolean Complement, self-checking embedded control schema, code ‘2 out of 4’, self-checking

I. INTRODUCTION

For the reason of the functional options advance, including sophistication of micro electronic gadgets, the opportunity of malfunction is just an augment in the event of various physical hindrances [1]. Based on statistics, prevalent

quantity of failures are within combinational circuits of automation sector [2]. Consequently, the vital task of nowadays is the synthesis of robust failsafe joint circuits as well as testing functional diagnostic systems arrangement [3 – 6].

Important approach of diagnostics combinational circuits array is considered effective methods of concurrent checking [7 – 9], which includes the arrangement of technical diagnostic procedures without any major function disconnection of those sites under supervision. The aforesaid is actually matters for critical technical sites with essential industrial performances [10].

Methods of concurrent checking considered abundance application to structure of initial combinational circuit, which may be as modified initial kit as well as some structure under survey with additional external schemas of monitoring or concurrent checking schema without any upgrade fulfillment [11]. Conventional option is a Back-up Method, during which preliminary combination is being equipped with self-copy and bit-to-bit comparison is being conducted while final unit service [12, 13]. Alternative approaches regarding functional control systems installation is reckoned the surveillance per being formed combined vectors of digital words of beforehand chosen noise-suppressing code [14 – 19] either supervision of realized functions regarding designated types, for instance, to self-dual functions [20 – 23].

Present essay is contributed to concurrent checking method being designed by authors, fulfilled in accordance with control of combinational circuits per two features – outputs vectors belongings to constant-weight code with function of digital words attachment to self-dual one of logic algebra.

II. STRUCTURAL SCHEMA OF CED SYSTEM

During combinational circuits supervision per two characteristics, the most effective one is the diagnostic option based on Boolean Complement Method performance [24].

The above method, considered function f_i , adjustment, via site under control into special one h_i by means of addition elements with module two (XOR elements) and control function g_i . Each function transformation should be completed in accordance with formula:

$$h_i = f_i \oplus g_i, \quad i = \overline{1, m}. \quad (1)$$

Values of function, being used during transformation, by means of combinational circuit of the formula (1), considered as universal and help us to form 'new' coded vector for afterward supervision performance upon designated feature. For example, in [25 – 29] we may see described options of control functions g_i calculations for digital words of constant-weight code arrangement, and in [30] – formation of self-dual functions. Well known structural schemas of 'self-dual parity' [20] and 'self-dual back-up' [22], which composed of original methods of supervision via parity with back-up and control by means of self-dual function. New method composed of work function f_i transformation into such functions h_i , which should be self-dual and at the same time the entire coded vector must be part of previously designated constant-weight code.

Statement 1. Supervision performance by features of self-dual with attachment of digital words to constant-weight codes is possible in the event of constant-weight codes ' r out of $2r$ ' (r – weight value of digital word of constant-weight codes), or $r/2r$ -code implementation only.

Fairness of the Statement 1 may be proved by the following: Initially, each vector $\langle f_1 f_2 \dots f_m \rangle$ should be transformed into vector $\langle h_1 h_2 \dots h_m \rangle$, with permanent weight. Then, on opposite input ports of set $\langle x_1 x_2 \dots x_r \rangle$ values of the entire bits of coded words $\langle h_1 h_2 \dots h_m \rangle$ must be inverted. Consequently, coded vector $\langle h_1 h_2 \dots h_m \rangle$ should compose equal quantity of on-bits with off-bits, otherwise the self-dual feature per each bit should not be ensured. Thus, the base for CED system can be only $r/2r$ -code, $2/4$ -code, $3/6$ -code, $4/8$ -code etc.

Important factor of this file is $2/4$ -code, which checker ($2/4$ -TSC) has simple structural schema composed of constant-weight code checker codes (Fig. 1). Besides, it requires each of combination out of matrix $\{0011; 1100; 0110; 1001\}$ [29] being presented toward input ports to ensure proper self-checking mode, which is considered minimum essential quantity to inspect checkers of constant-weight codes.

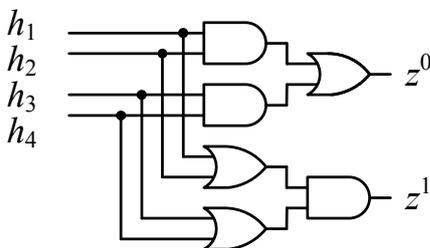


Fig. 1. Structural schema $2/4$ -TSC.

Let us consider each function h_i should be self-dual, consequently for the purpose of supervision, self-dual checker of the function SSC (self-checking self-dual checker) should be applied, See Fig. 2. Self-dual signal f^* by means of resistance line, which equals single cycle time of impulse subsequence a , should be transformed into two-rail

signal $\langle v_1 v_2 \rangle$. SSC provided with two outputs ports and while self-dual input performance signal ought to form two-rail signal $\langle 01 \rangle$ either $\langle 10 \rangle$ per exit ports. In the event of self-dual input indication fault, non-two-rail signal must be formed on the output port.

Outputs of checkers SSC and $2/4$ -TSC should be unified on inputs ports of self-checking comparator, being designed based on compressed moduli of two-rail signal TRC (two-rail checker) [31] (Fig. 3).

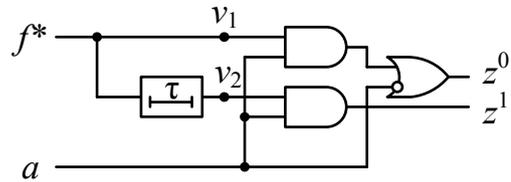


Fig. 2. Structural schema SSC.

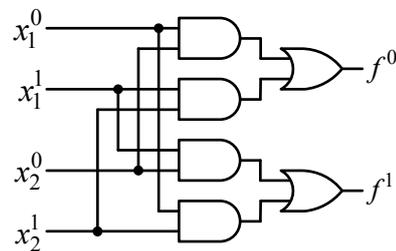


Fig. 3. Structural schema TRC.

Alternative option is comprised of preliminary implementation compression circuit of self-dual signals per function h_i with afterward monitoring of self-dual signal.

Combined structural schema of concurrent checking based on pair of characteristics is presented in Fig. 4.

Based on the abovementioned consideration we may state the following:

Statement 2. Within CED system based on Boolean Complement up to constant-weight code with self-dual functions, any error should be identified into coded word $\langle h_1 h_2 h_3 h_4 \rangle$, except the one, which does not violates designated weight and at the same time obtainable in the form of same distortion on the opposite sides of the input port sets.

Evidently, the identification feature, compared to supervision per single characteristic, is much better. Structural schema, shown in Fig. 4, considered as a basic one. In case of monitoring arrangement regarding multi-outputs combinational circuits, those outputs should be divided into sub-groups per four exits (sub-settings may intersect). Per each of outputs sub-settings separate control circuit should synthesize. Outputs of separate supervision schemas must be unified per outputs of self-checking comparator.

III. OPTIONS OF COMPLEMENT FUNCTIONS CALCULATION

Let us take a look at logical statements completion with description of control function g_i , in accordance with subsequent additional definition of the aforesaid values.

We may focus on designated task fulfillment by means of combinational circuit example in accordance with true table (Table 1).

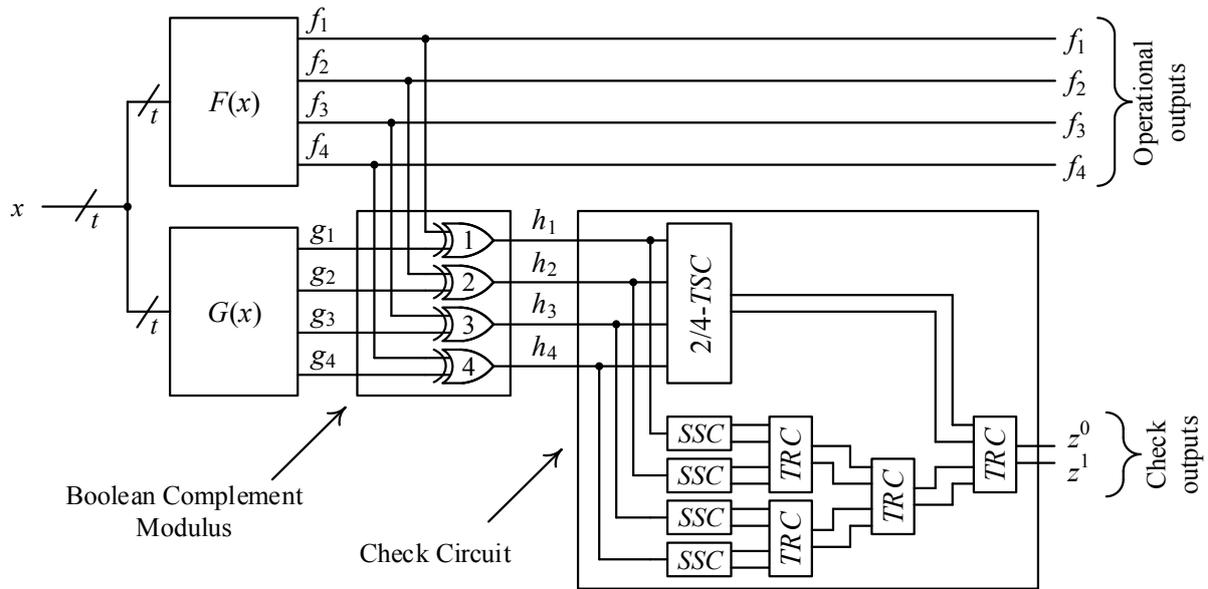


Fig. 4. Structural schema of CED system.

We shall receive check functions values g_1, g_2, g_3, g_4 , which may help us to issue proper supervision of designated combinational circuit in accordance with proposed method.

TABLE I. TRUE TABLE OF COMBINATIONAL CIRCUIT

No.	x_1	x_2	x_3	x_4	f_1	f_2	f_3	f_4
0	0	0	0	0	0	1	0	1
1	0	0	0	1	1	1	1	1
2	0	0	1	0	1	1	1	1
3	0	0	1	1	1	0	0	0
4	0	1	0	0	1	0	1	1
5	0	1	0	1	0	0	0	0
6	0	1	1	0	1	1	0	1
7	0	1	1	1	0	1	0	1
8	1	0	0	0	1	0	1	1
9	1	0	0	1	0	0	1	0
10	1	0	1	0	0	1	1	1
11	1	0	1	1	0	1	0	1
12	1	1	0	0	1	0	0	1
13	1	1	0	1	0	0	1	0
14	1	1	1	0	0	1	1	1
15	1	1	1	1	1	0	0	0

Statement 3. For the purpose of CED system based on two characteristics would get self-checking feature per single permanent errors, the following conditions must be completed:

1. Each coded word $\langle h_1 h_2 h_3 h_4 \rangle$ should have weight $r=2$;
2. At least single testing combination ought to be arranged per 2/4-TSC: {0011; 1100; 0110; 1001};
3. Minimum one testing combination for each element XOR: {00; 01; 10; 11} must be formed;
4. Values of each function h_1, h_2, h_3, h_4 must be inverted on opposite input sets.

The easiest way for the above terms achievement is defined within Statement 3 with the following algorithm employment:

Algorithm 1. Additional definition of controlled values, including the condition of circuit self-checking feature with afterward inspection of Boolean Complement segment:

1. The presence of at least two off-bits with two on-bits of each function f_1, f_2, f_3, f_4 per total input sets should be the matter of control.
2. Into columns of the first (either second) half of the Table regarding the middle part, matching those coded words $\langle h_1 h_2 h_3 h_4 \rangle$, in random way, but equable, we shall insert combinations $\langle 0011 \rangle$ & $\langle 0110 \rangle$ or $\langle 0011 \rangle$ and $\langle 1001 \rangle$, or $\langle 1100 \rangle$ and $\langle 0110 \rangle$, or $\langle 1100 \rangle$ and $\langle 1001 \rangle$;
3. By means of inverse values filling on opposite sets reckoning Table half, columns of the rest of the Table must be filled up according to coded words $\langle h_1 h_2 h_3 h_4 \rangle$.
4. Values of control functions $g_i = f_i \oplus h_i, i = \overline{1,4}$ should be calculated;

5. Being formed combination via element of addition per module two within Boolean Complement segment are being defined;

6. Based on Item 2 with Item 3, requirements concerning testing matrix formation per 2/4-TSC with functions h_1, h_2, h_3, h_4 of self-dual feature will be fulfilled. It is essential to check the formation of the entire matrix of testing combination for elements of addition via modulus two in accordance with Boolean Complement segment.

7. In case of for some element of addition of modulus two per Boolean Complement segment one cannot arrange the required testing combination, then there is a need to alter row filling procedure per the stage of Item 2 completion.

In accordance with the abovementioned we shall receive Table 2.

Algorithm 1 is suitable to apply in the event of multiple amounts of input disposal variables, for the reason that possibility of result achievement concerning formation of testing combination of modulus two per Boolean Complement

segment in the event of input parameters augmentation is raising.

Insofar as is concerned alternative for the Algorithm 1, we may state the application of the option of prior checking performance regarding elements per modulus two.

Algorithm 2. Additional definition of controlled function values, including the condition of controllability:

1. The matter of control must be at least the presence of two off-bits with two on-bits per each function f_1, f_2, f_3, f_4 on the entire input sets;

2. Additional definition for functions h_1, h_2, h_3, h_4 should be completed, including the necessity of testing combination arrangement by modulus two within Boolean Complement segment:

– for at least a single event $f_i=0$ function value h_i must be additionally defined 0 and for at least one case where $f_i=0$ – equals 1;

– for at least a single event $f_i=1$ function value h_i must be additionally defined 0 and for at least one case where $f_i=0$ – equals 1;

3. Hence, if the additional definition h_i for the case of the same inputs sets resulted in pop up of non-coded word of 2/4-code, another one definition should take place, but if none of those definition options complies with designated term, it means that self-checking device based on the above approach cannot be fulfilled;

4. Function values h_1, h_2, h_3 ought to be filled up and h_4 in lines, being located symmetrically regarding table center: filled up digits ought to be inverted and to be placed into designated lines with symmetrical disposition;

5. Columns of coded words should be loaded up $\langle h_1 h_2 h_3 h_4 \rangle$ for those lines, where values were placed previously;

6. Existence of four 2/4-TSC testing combination should be checked up;

7. In accordance with conditions reckoning coded words of 2/4-code formation for 2/4-TSC with self-dual functions h_1, h_2, h_3, h_4 arrangement, the rest of columns should be filled up.

8. The abovementioned algorithms allow us to realize CED systems of combinational circuits per two characteristics.

IV. PRACTICAL IMPLEMENTATION

For the research purposes relating to errors identification possibilities into present article, experiments within structural schema of concurrent checking were completed via *Multisim* modeling of single permanent faults performance for the aforementioned example.

Realization process appeared to be not simplistic one for the reason that via conventional instruments of *Multisim*, signal lag performance was not easy to accomplish by means of checker self-dual functions. The solution was taken concerning two JK-triggers application, which are functioning like D-triggers (U15A with U9A). For example, of self-dual analog checker fulfilled within *Multisim*, modeling environment, refer to Fig. 5.

Impulse generator with frequency of 10 Hz with some phase difference is being connected to inputs of synchronized triggers. Moreover, to one of triggers generator is linked via U16D-inverter, thus triggers should come to action in different time sector. It ought to be pointed out that checker is linked to one more impulse generator with frequency 10 Hz (U13).

TABLE II. SIGNALS ON LINES OF CED SCHEMA

No.	Values of work functions				Values of check functions				Coded words on exit of Boolean Complement segment				Formation of testing combination for Addition elements via modulus two			
	f_1	f_2	f_3	f_4	g_1	g_2	g_3	g_4	h_1	h_2	h_3	h_4	XOR_1	XOR_2	XOR_3	XOR_4
0	0	1	0	1	0	0	1	1	0	1	1	0	00	10	01	11
1	1	1	1	1	1	0	0	1	0	1	1	0	11	10	10	11
2	1	1	1	1	1	0	0	1	0	1	1	0	11	10	10	11
3	1	0	0	0	1	1	1	0	0	1	1	0	11	01	01	00
4	1	0	1	1	1	0	0	0	0	0	1	1	11	00	10	10
5	0	0	0	0	0	0	1	1	0	0	1	1	00	00	01	01
6	1	1	0	1	1	1	1	0	0	0	1	1	11	11	01	10
7	0	1	0	1	0	1	1	0	0	0	1	1	00	11	01	10
8	1	0	1	1	0	1	1	1	1	1	0	0	10	01	11	11
9	0	0	1	0	1	1	1	0	1	1	0	0	11	01	11	00
10	0	1	1	1	1	0	1	1	1	1	0	0	01	10	11	11
11	0	1	0	1	1	0	0	1	1	1	0	0	11	10	00	1
12	1	0	0	1	0	0	0	0	1	0	0	1	00	00	00	10
13	0	0	1	0	1	0	1	1	1	0	0	1	11	00	11	01
14	0	1	1	1	1	1	1	0	1	0	0	1	01	11	11	10
15	1	0	0	0	0	0	0	1	1	0	0	1	10	00	00	01

As is shown in Fig. 5, in the event of on-bit on the output of U13-generator, X1 and X2 light bulbs should inform us regarding the signal from triggers output. In case of off-bit output signal from U13-generator, we shall receive light bulbs signal 10. Consequently, the checker is being formed two-rail signal in the event of off-bit signal with output or when the output signal is on-bit, but triggers state is considered the opposite one. For the further explanation let us take a look at simplified circuit under control of designated checker in Fig. 6.

Comparison in-coming signal with previous one is being conducted via self-dual checker on its input port. For this reason within inputs of under supervision schema elements of addition U1C, U1A, U1B per module two are being installed. One port of those elements is linked to designated signal source and another one is connected to U10-impulse generator with impulse frequency two times less then checker generator. As has been mentioned, checker generators have some phase difference, but without the above difference the moment of force will appear, when total generators should simultaneously be activated, which due to underdetermined condition. The aforesaid condition resulted in false error signal on checker output. Consequently, at the moment when U10 generator issued off-bit signal, opposite designated sets should be presented per outputs schema ports. Next, if the circuit is fulfilled as self-dual one, then self-dual signal must be issued within output port. For the further description let us have a look at temporary circuit function diagram in Fig. 6 (Fig. 7).

In Fig. 7 the following letters A – H are chosen per lines of diagrams located in Fig. 6. During starting moment within

A-generator output on-bit signal ought to be identified, which means opposite input set is being conducted toward circuit inputs. Then on output of circuit under supervision we shall see on-bit signal and at the same time within checker outputs two-rail signal should appear, for the reason that on outputs of B&C generators the off-bit event exists. Consequently, B&C generators must alter self-conditions from off-bit to on-bit. At this particular moment output signal of under control circuit must be recorded by F-trigger. On checker outputs two-rail signal should be formed, due to the fact that different signals are presented on triggers outputs. Hereafter, B&C generators should alter self-status from on-bit into off-bit and immediately output circuit condition must be recorded via F-trigger. Both triggers have the same status quo at this moment, but within checker output two-rail signal should be detected, because C-generator must work out off-bit signal. Then G-generator ought to change self-state from on-bit unto off-bit, which means that proper signal should arrive on schema inputs and at this moment opposite signal to previous one should be defined (N.B. for the term that the system is self-dual!). After short period of time this signal must be registered via E-trigger and both triggers should be in proper conditions, which resulted in two-rail signal on checker output. Then, after status change of B-generator, information from output of the system under control must be recorded into F-trigger. Based on the above, triggers are being adjusted in the same condition, but in the event of C-generator has off-bit within output port, checker should issue two-rail signal. Further F-generator ought to switch into on-bit condition and the cycle must be repeated one.

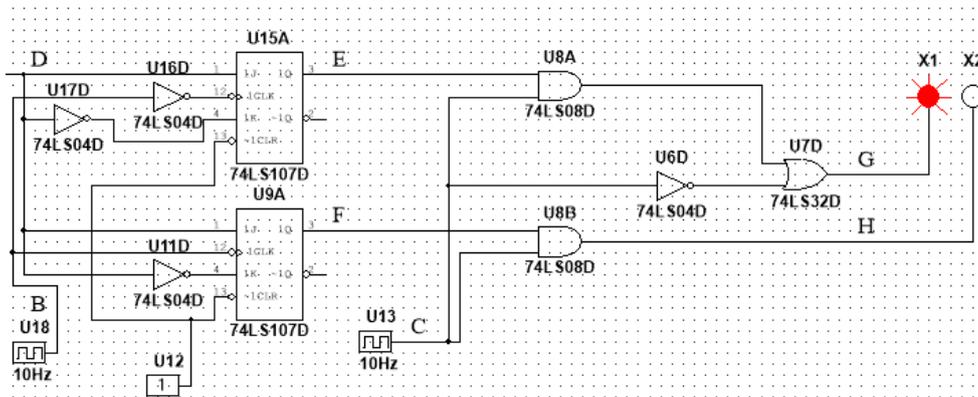


Fig. 5. Self-dual checker.

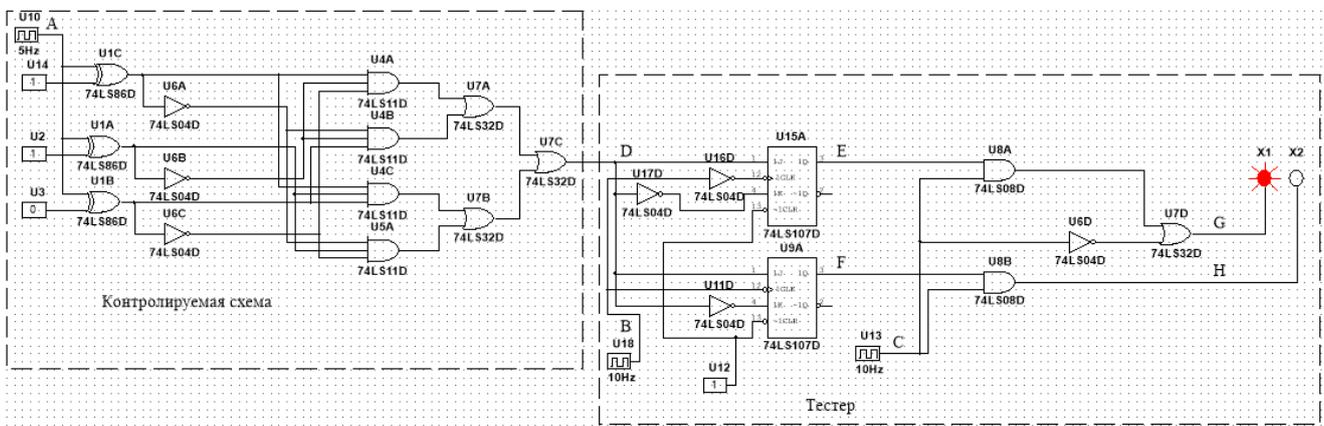


Fig. 6. Example of self-dual checker implementation.

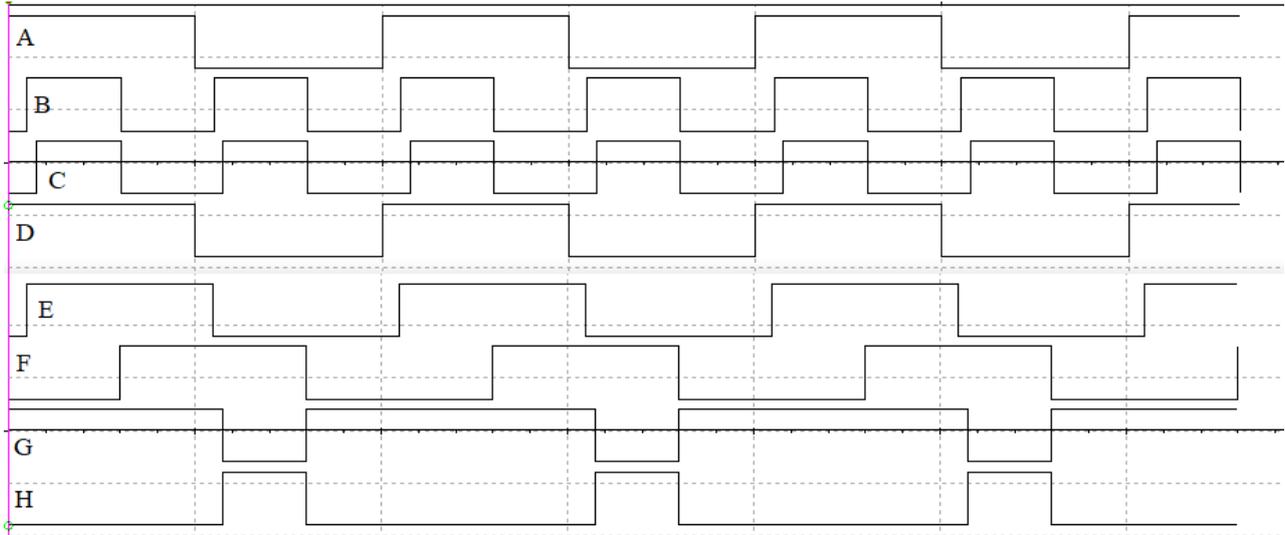


Fig. 7. Temporary diagram of schema functioning.

Consequently, if the circuit under control is completed as a self-dual, which means, it must alter output status on the opposite one in the event of A-generator output signal was changed on opposite one. In case of opposite signals will appear on circuit's output one-by-one, so triggers should be within visa-versa status in the event of C-generator has on-bit on the exit port. As a result of opposite signals from triggers will be delivered to outputs of checker and consequently two-rail signal will be fulfilled. Two-rail signal will appear in the same way at the moment when off-bit event will be issued on C-generator output.

For the option of non-self-dual circuit, the same signal will be presented within input port until the event of input alteration. In this case, triggers ought to be installed in the same conditions and checker should work out non-two-rail signal in case of C-generator state on-bit on exit port.

It should be mentioned that per option of U13-generator arrangement the same way as U18-generator, we should complete the switch moment to on-bit status on schemas outputs signals from triggers presentation and as for U15A-trigger, checker input data will be recorded. Hence, trigger status quo is being changed gradually, thus there ought to be short period of time when false error signal must be identified within checker outputs. For the above reason U13-generator has bigger difference compared to U18. In this case into off-bit condition U18-generator is being switched much faster and as a result, we shall receive the moment when U9A-trigger is being recorded the signal from checker exit and both triggers are being switched into the equal condition. At the same time on the exit of U13-generator on-bit is fixed and signals from triggers are being conducted per outputs of checker, consequently, false signal of error will arise. To avoid this trouble, we need U13-generator to change self-status into off-bit condition prior either at the same time with U-19 unit. To achieve this goal, we must reduce duty cycle of U13-generator.

Presented self-dual checker layout in Fig. 8 is not suitable per actual circuits application, for the motive of embedded memory elements A. Hence, in our case (during modeling performance) it does not matter, inasmuch as we need just data regarding detection feature presence.

Testing procedure of being modeling circuit was conducted in an analogous way per each h -function. To each checker 10 Hz generator with little phase difference was connected. Further, outputs of checkers should be unified by means of *TRC* schemas. Innovative schema of checker must be linked with constant-weight code '2 out of 4' checker as well via *TRC*. Each schema input is being equipped with 'summation via module 2' elements, one of enter ports of which must be linked to generator of inputs signals and the other one to 5 Hz impulse generator. As can be seen from the above, in the event of at least single h -function will not be self-dual, this way non-two-rail signal of 10 Hz must be worked out.

For simulation purpose layout of Table 1 was implemented within *Multisim* environment. As a consequence of functions minimization via true tables, we resume the following:

$$\begin{aligned}
 f_1 &= \overline{x_1 x_2 x_4} \vee \overline{x_1 x_3 x_4} \vee \overline{x_2 x_3 x_4} \vee \overline{x_1 x_3 x_4} \vee \overline{x_1 x_2 x_3 x_4}; \\
 f_2 &= \overline{x_3 x_4} \vee \overline{x_1 x_2 x_3} \vee \overline{x_1 x_2 x_3} \vee \overline{x_1 x_2 x_3}; \\
 f_3 &= \overline{x_2 x_3 x_4} \vee \overline{x_1 x_2 x_3} \vee \overline{x_1 x_3 x_4} \vee \overline{x_2 x_3 x_4} \vee \overline{x_1 x_3 x_4} \vee \overline{x_1 x_2 x_3 x_4}; \\
 f_4 &= \overline{x_1 x_4} \vee \overline{x_1 x_2 x_3} \vee \overline{x_1 x_2 x_3} \vee \overline{x_1 x_2 x_3} \vee \overline{x_1 x_4}.
 \end{aligned}$$

The above formulas are being described those functions of schema being realized via two level disjunctive form of presentation. For aliquot errors imitation formula circuit results were transformed as follows:

$$\begin{aligned}
 f_1 &= yx_4 \vee \overline{x_1 x_3 x_4} \vee \overline{x_2 x_3 x_4} \vee zx_3 \vee ak; \\
 f_2 &= p \vee \overline{x_1 x_2 x_3} \vee \overline{x_1 k} \vee \overline{y x_3}; \\
 f_3 &= bx_4 \vee \overline{x_1 b} \vee \overline{x_1 p} \vee \overline{x_2 p} \vee \overline{ax_3} \vee \overline{zx_2 x_3}; \\
 f_4 &= z \vee \overline{x_1 k} \vee \overline{x_1 x_2 x_3} \vee \overline{y x_3} \vee \overline{x_1 x_4}; \\
 y &= \overline{x_1 x_2}; \\
 z &= \overline{x_1 x_4}; \\
 k &= \overline{x_2 x_3}; \\
 a &= \overline{x_1 x_4}; \\
 b &= \overline{x_2 x_3}.
 \end{aligned}$$

Combinational circuit was issued in accordance with end formula together with CED system by means of our elaborated within present article *Multisim Option*. To being implemented elements per several functions (y, z, k, a), step-by-step permanent malfunctions were gradually inserted. It was completed via designated element disconnection and further exchange via the source of logical signals. Next stage was successive delivery per various input sets. Outcome values of exit ports were the matter of comparison being presented within digits of Table 1. In the event of registration values difference we had the error factor into data vector. In case of two-rail signal checker evidence, the error was classified as non-detectable one.

Actually per eight malfunctions of schema we received 45 errors including 16 – multiple-bit errors and the rest were single-bit errors. The entire faults were identified within CED system.

It is worth noting that during such option of supervision, checker did show the signal of an error even per case of absolute error absence with exit ports, but the fault was evident. It was the matter of fact, that, in case of malfunction existence, it was not defined within selected set, but was identified in the opposite one. As a consequence of the aforementioned, self-dual feature of the one of functions was disturbed and on the exit of checker the error signal was recorded.

V. CONCLUSION

Method of synthesis of CED system per two characteristics (coded words attachment to designated code set with functions of self-dual quality of logic algebra) being discussed in our article, allows us to upgrade detection quality of diagnostic system compared to supervision based on single characteristic.

Completed research has shown that one of the cost effective options regarding the arrangement of CED systems per 2 features, is considered comprised implementation of self-dual complement with monitoring per code ' r out of $2r$ ' ($r/2r$ -code), where r – weight of coded vector. We believe, more feasible is application of $2/4$ -code for the required task fulfillment, which checker has simple structure for the complete examination of testing variation.

Resume of modeling performance reckoning the CED system in the event of single malfunction insertion within exit of inner logical element considered as a proof of monitoring performance effectiveness per two characteristics.

Being described method of synthesis concerning the CED system, we believe, is the advanced one, with obtainable results is feasible to assume during the design procedure of combinational circuit concurrent checking based on elements of modern software.

REFERENCES

- [1] R. Ubar, J. Raik, and H.-T. Vierhaus "Design and Test Technology for Dependable Systems-on-Chip (Premier Reference Source)", Information Science Reference, Hershey – New York, IGI Global, 2011, 578 p.
- [2] A. Stempkovskiy, D. Telpukhov, and V. Nadolenko "Development of Resynthesis Flow for Improving Logical Masking Features of Combinational Circuits", Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS'2018), Kazan, Russia, September 14-17, 2018, pp. 34-40, doi: 10.1109/EWDTS.2018.8524629.
- [3] P.K. Lala "Principles of Modern Digital Design", New Jersey: John Wiley & Sons, 2007, 419 p.
- [4] M. Goessel, V. Ocheretny, E. Sogomonyan, and D. Marienfeld "New Methods of Concurrent Checking: Edition 1", Dordrecht: Springer Science+Business Media B.V., 2008, 184 p.
- [5] Z. Navabi "Digital System Test and Testable Design: Using HDL Models and Architectures", Springer Science+Business Media, LLC 2011, 435 p.
- [6] V. Hahanov "Cyber Physical Computing for IoT-driven Services", New York, Springer International Publishing AG, 2018, 279 p., doi: 10.1007/978-3-319-54825-8.
- [7] M. Nicolaidis, and Y. Zorian "On-Line Testing for VLSI – A Compendium of Approaches", Journal of Electronic Testing: Theory and Application (JETTA), 1998, vol. 12, issue 1-2, pp. 7-20, doi: 10.1023/A:1008244815697.
- [8] S. Mitra, and E.J. McCluskey "Which Concurrent Error Detection Scheme to Choose?", Proceedings of International Test Conference, 2000, USA, Atlantic City, NJ, 03-05 October 2000, pp. 985-994, doi: 10.1109/TEST.2000.894311.
- [9] K. Mohanram, C.V. Krishna, and N.A. Touba "A Methodology for Automated Insertion of Concurrent Error Detection Hardware in Synthesizable Verilog RTL", IEEE International Symposium on Circuits and Systems (ISCAS 2002), 26-29 May 2002, vol. 1, pp. 577-580, doi: 10.1109/ISCAS.2002.1009906.
- [10] V. Kharchenko, Yu. Kondratenko, and J. Kacprzyk "Green IT Engineering: Concepts, Models, Complex Systems Architectures", Springer Book series "Studies in Systems, Decision and Control", vol. 74, 2017, 305 p, doi: 10.1007/978-3-319-44162-7.
- [11] D.V. Efanov, V.V. Sapozhnikov, and V.I. Sapozhnikov "Synthesis of Self-Checking Combinational Devices Based on Allocating Special Groups of Outputs", Automation and Remote Control, 2018, issue 9, pp. 1607-1618, doi: 10.1134/S0005117918090060.
- [12] J. Borecký, M. Kohlík, and H. Kubátová "Parity Driven Reconfigurable Duplex System", Microprocessors and Microsystems, 2017, Vol. 52, pp. 251-260, doi: 10.1016/j.micpro.2017.06.015.
- [13] V.V. Sapozhnikov, V.I. Sapozhnikov, D.V. Efanov, and Dmitriev "New Structures of the Concurrent Error Detection Systems for Logic Circuits", Automation and Remote Control, 2017, vol. 78, issue 2, pp. 300-313, doi: 10.1134/S0005117917020096.
- [14] D. Das, and N. A. Touba "Synthesis of Circuits with Low-Cost Concurrent Error Detection Based on Bose-Lin Codes", Journal of Electronic Testing: Theory and Applications, 1999, vol. 15, issue 1-2, pp. 145-155, doi: 10.1023/A:1008344603814.
- [15] A.Yu. Matrosova, I. Levin, and S.A. Ostanin "Self-Checking Synchronous FSM Network Design with Low Overhead", VLSI Design, 2000, vol. 11, is. 1, pp. 47-58.
- [16] E. Fujiwara "Code Design for Dependable Systems: Theory and Practical Applications", John Wiley & Sons, 2006, 720 p.
- [17] S. Ostanin "Self-Checking Synchronous FSM Network Design for Path Delay Faults", Proc. of 15th IEEE EWDTS'2017, Novi Sad, Serbia, September 29 – October 2, 2017, pp. 696-699.
- [18] G. Tshagharyan, G. Harutyunyan, S. Shoukourian, and Y. Zorian "Experimental Study on Hamming and Hsiao Codes in the Context of Embedded Applications", Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 25-28, doi: 10.1109/EWDTS.2017.8110065.
- [19] A. Stempkovskiy, D. Telpukhov, S. Gurov, T. Zhukova, and A. Demeneva "R-Code for Concurrent Error Detection and Correction in the Logic Circuits", 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIconRus), 29 January – 1 February 2018, Moscow, Russia, pp. 1430-1433, doi: 10.1109/EIconRus.2018.8317365.
- [20] V.I. Sapozhnikov, A. Dmitriev, M. Goessel, and V.V. Sapozhnikov "Self-Dual Parity Checking – a New Method for on Line Testing", Proceedings of 14th IEEE VLSI Test Symposium, USA, Princeton, 1996, pp. 162-168.
- [21] V.I. Sapozhnikov, V. Moshanin, Sapozhnikov V.V., and M. Goessel "Self-Dual Multi-Output Combinational Circuits with Output Data Compaction", Compendium of Papers IEEE European Test Workshop (ETW'97), Cagliari, Italy, May 28 – 30, 1997, pp. 107-111.
- [22] V.I. Sapozhnikov, V.V. Sapozhnikov, A. Dmitriev, and M. Goessel "Self-Dual Duplication for Error Detection", Proceedings of 7th Asian Test Symposium, Singapore, 1998, pp. 296-300.

- [23] V.I. Sapozhnikov, Moshanin V., Sapozhnikov V.V., and M. Goessel "Experimental Results for Self-Dual Multi-Output Combinational Circuits", *Journal of Electronic Testing: Theory and Applications*, 1999, Vol. 14, Issue 3, pp. 295-300, doi: 10.1023/A:1008370405607.
- [24] M. Goessel, A.V. Morozov, V.V. Sapozhnikov, and V.I. Sapozhnikov "Logic Complement, a New Method of Checking the Combinational Circuits", *Automation and Remote Control*, 2003, vol. 1, issue 1, pp. 153-161.
- [25] A. Morozov, M. Gössel, V. Sapozhnikov, and V.I. Sapozhnikov "Complementary Circuits for On-Line Detection for 1-out-of-3 Codes", *ARCS 2004 – Organic and Pervasive Computing, Workshops Proceedings*, March 26, 2004, Augsburg, Germany, pp. 76-83.
- [26] M. Goessel, A.V. Morozov, V.V. Sapozhnikov, and V.I. Sapozhnikov "Checking Combinational Circuits by the Method of Logic Complement", *Automation and Remote Control*, 2005, vol. 66, issue 8, pp. 1336-1346.
- [27] S. Halder, S.S. Roy, and S.K. Sen "An Optimized Concurrent Self-Checker using Constraint-Don't Cares and 1-out-of-4 Code", *National Conference (AECDISC-2010) in Asansol Engineering College*, 1-2 August 2010.
- [28] D.K. Das, S.S. Roy, A. Dmitriev, A. Morozov, and M. Gössel "Constraint Don't Cares for Optimizing Designs for Concurrent Checking by 1-out-of-3 Codes", *Proceedings of the 10th International Workshops on Boolean Problems, Freiberg, Germany, September, 2012*, pp. 33-40.
- [29] V. Sapozhnikov, V.I. Sapozhnikov, and D. Efanov "Concurrent Error Detection of Combinational Circuits by the Method of Boolean Complement on the Base of «2-out-of-4» Code", *Proceedings of 14th IEEE East-West Design & Test Symposium (EWDTS'2016)*, Yerevan, Armenia, October 14-17, 2016, pp. 126-133, doi: 10.1109/EWDTS.2016.7807677.
- [30] M. Goessel, A.V. Dmitriev, V.V. Sapozhnikov, and V.I. Sapozhnikov "A Functional Fault-Detection Self-Test for Combinational Circuits", *Automation and Remote Control*, 1999, Vol. 60, Issue 11, pp. 1653-1663.
- [31] D. Nikolos "Self-Testing Embedded Two-Rail Checkers", *Journal of Electronic Testing: Theory and Applications*, 1998, Vol. 12, Issue 1-2, pp. 69-79, doi: 10.1023/A:1008281822966.

Decision making in VLSI components placement problem based on grey wolf optimization

Elmar V. Kuliev¹, Vladimir VI. Kureichik, Ilona O. Kursitys
Southern Federal University
Rostov-on-Don, Russia
¹ i.kursitys@mail.ru

Abstract— This article is devoted to one of the key automated design tasks, which is the problem of the VLSI components placement. The possibilities of bioinspired algorithms development and application for the purpose of the effective decision support in the CAD sphere are of a great interest today. With that, there is a constant conflict between the CAD complexity and requirements of the effective decision support in real time. Such methods as parallelizing of the decision support process, increasing the number of operators or users are not able to solve the tasks mentioned above completely. The paper considers the new technologies combining computer science, bionics and computer-aided design as the way to solve this problem. In this context, the development of the new principles and methods of effective decision support in design and management problems is of a great economic and social importance and is claimed to be relevant today. The paper describes the living nature algorithm based on the grey wolf pack behavior. The authors formulate the problem of the VLSI components placement on a set of positions of a discrete work field. The modified technology of the bioinspired algorithms development and the main steps of the grey wolf pack behavior in terms of the VLSI placement problem are presented in the paper. The computational experiments show the results of the developed approach. The main purpose of the research is to estimate the possibilities of integrated nature-inspired methods for the purpose of CAD problems solving based on the grey wolf pack behavior in wildlife.

Keywords— *swarm intelligence, genetic algorithm, objective function, neighborhood, wolf pack.*

I INTRODUCTION

Designing complex objects requires heavy spending time and human resources, thus, it is reasonable to use different tools for computer aided design. In this context, computer is represented as a design tool and as the object of design. Design automation depends on quality of computing equipment and its hardware components, i.e. very large scale integration (VLSI) and very high speed integrated circuits (VHSIC). Computer aided design requires the following types of support [1-6,8,11]:

1. Mathematical support – transition from the object description to its mathematical model and mathematical designing, i.e. methods and algorithms of solving such tasks.

2. Software support – a set of program solutions implementing operations and procedures necessary to obtain description of the designed object.

3. Information support – structured data used for reference and design material.

4. Linguistic support – terminology, programming and description languages.

One of the most popular tasks related to computer aided design (CAD) is the VLSI floor planning. It involves placing blocks on the crystal's field. The blocks are obtained after the partitioning stage, have an assigned area and no fixed dimensions. The task involves two simultaneous tasks: elements placements and fixing the dimensions of each block. The placement problem includes defining a place for each element and block on the crystal's field. Following characteristics are considered: length of connections, allocation of blocks among the crystal, time delay, crystal's area and dimensions. The main purpose of placement is to provide the best conditions for the further routing. Criteria and estimations are introduced and optimized in terms of the best placement conditions.

II PROBLEM STATEMENT

The VLSI placement problem can be formulated as follows. We are given a set of elements with contacts (outputs), elementary connections relating the elements contacts and a switchboard to place the elements. The task is to place the elements on the switchboard in terms of the quality criteria optimization. The main purpose is to minimize the common area of the crystal, create the routing conditions and minimize the common total length of connections ($L(G) \rightarrow \min$) [3,7,9,10]. In general terms, the placement task can be formulated as follows: the field of the installation space is divided into a set of positions which quantity is greater or equals to number of placed elements. Each element can take only one seat, and the distance between the seats is described by a symmetric distance matrix. A set of elements interconnected by a set of electric circuits is to be mapped into a set of P to provide the extremum of the placement quality objective function. The placement task is reduced to mapping the assigned graph into a grid in such way that the vertexes are placed in the grid points and the total length is minimal for possible ways of matching the graph vertexes with grid points. The original area is covered with a rectangular

This research is supported by the Council for Grants (under RF President), the project # MK-1480.2018.9

coordinate system with s and t axes. Each cell of the area is placed with a circuit element.

III GREY WOLF OPTIMIZATION ALGORITHM

In recent times bioinspired methods are popular to use in solving the computer aided design (CAD) tasks of placement and designing. Bioinspired methods are based on using algorithms describing the natural evolutionary processes. Genetic algorithms are also referred to bioinspired methods. Such approach allows us to obtain optimal and quasioptimal solutions in polynomial time.

The paper considers the grey wolf algorithm which was suggested in 2014. It is based on the mechanisms of grey wolf hierarchy and hunting process in natural life. There are four types of wolves: Alpha, Beta, Delta and Omega, which are used for modeling of the leadership hierarchy. The following stages of hunting are implemented in the optimization process: tracking, encircling and attacking the prey. Grey wolves belong to the Canidae family and are at the top of the vermin hierarchy, i.e. at the top of the food chain. Grey wolves mostly live in packs consisting of 5-12 species. A very strict social dominance hierarchy is of a great interest.

The leaders are represented by male and female species referred as Alpha. The Alpha makes decisions on hunting, sleeping place, time to wake up, etc. However, there is a democratic behavior, when the Alpha follows other wolves of the pack. At a gathering Alpha is recognized by the whole pack by holding their tails down. The Alpha wolves are allowed to mate in the pack only. The Alpha is not always the strongest wolf of the pack, but the best in terms of the pack management. Thus, the organization and discipline in the pack are more important than its strength.

A Beta-wolf is the second in the hierarchy. Beta wolves are subordinate of Alpha and help them make decisions or in other activity of the pack. A Beta wolf can be a male or a female wolf and it is the best candidate to be the Alpha if one of the Alphas are old or dead. Beta wolf is to respect the Alpha and manage the wolves of lower levels. It is represented as the adviser to the Alpha and a discipliner for the pack. The Beta reinforces the Alpha's commands and gives the feedback to the Alpha [12,15].

Omega-wolf is the lowest in the hierarchy, Omega-wolves always obey all dominating wolves. It can seem to be that this type of wolves is not necessary, but in the case of their absence there are the problems and fighting in the pack. The other wolves release their aggression and violence on the Omegas, which help save the hierarchy structure.

The other type of wolves is called Delta. Delta-wolves obey the Alphas and Betas, but rank above the Omegas. This category includes: scouts, sentinels, elders, hunters and caretakers. The scouts are responsible for the boundaries watching and warning the pack if the danger comes. The sentinels protect the pack. The elders are the experiences wolves (Alphas and Betas in the past). Hunters help Alpha and Beta in hunting.

To build a mathematical model of the social hierarchy of wolves in terms of designing, let us consider the Alpha (α) as the fittest solution. Thus, the second and the third best solutions are

called Beta (β) and Delta (δ) respectively. All other solutions are considered as Omega (ω). In terms of the Grey Wolf Optimization (GWO) algorithm, α , β and δ are responsible for hunting (optimization).

Originally, the Grey Wolf Optimization (GWO) algorithm is targeted at solving the vector tasks, thus, its main principles should be adapted for solving the placement task. The initial population is to be represented as chromosomes, and encoded/decoded in accordance with the following rule: the number of genes in the chromosomes are seats, and the value of genes are numbers of elements placed at the seats.

As it was mentioned above, grey wolves encircle the prey while hunting. To model the behavior of the encircling wolves, let us introduce the following equations [3,13]:

$$D = |\vec{C} \times \vec{X}_p(t) - \vec{X}(t)|;$$

$$\vec{X}(t + 1) = \vec{X}_p(t) - \vec{A} \times \vec{D},$$

where t denotes the current iteration, \vec{A} and \vec{C} are coefficient vectors, \vec{X}_p – is the position of the ‘prey’ vector, and \vec{X} shows the position of the ‘grey wolf’ vector.

Vectors \vec{A} and \vec{C} are calculated as follows:

$$\vec{A} = 2\vec{a} \times \vec{r}_1 - \vec{a};$$

$$\vec{C} = 2 \times \vec{r}_2,$$

where \vec{a} is reduced from 2 to 0 linearly during the iterations, and r_1, r_2 are random vectors in [0,1]. In terms of the stated task, parameter a is represented as the one-point mutation among all the species, and is reduced uniformly during all iterations. Random vectors r_1, r_2 allow the wolves to get to any position between two concrete points. In terms of the n-dimensional space, grey wolves move around the best solution in a hypercube (Fig.1).

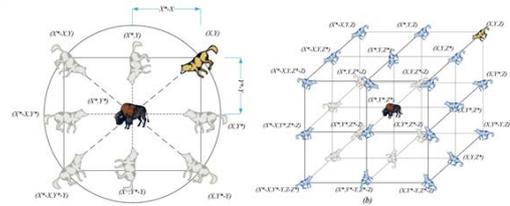


Fig. 1. Updating the position of wolves

Considering the equations (3) and (4) mentioned above, being in the position (X, Y), grey wolf can update its position in accordance with the prey's position (X*, Y*). Different positions around the best agent can be obtained in relation to the current position by adjusting the values of vectors \vec{A} и \vec{C} . For instance, (X*-X, Y*) can be obtained by setting $\vec{A} = (1,0)$ and $\vec{C} = (1,1)$.

Grey wolf can recognize the prey's position and encircle it. Hunting is usually implemented by the Alpha. Sometimes Beta and Delta can participate in hunting, too. However, in terms of the abstract space, we do not have information about position of the prey (optimum). To model the hunting behavior of grey wolves, let us assume that the Alpha (candidate for the best solution), Beta and Delta see the potential position of the prey more clearly. Thus, we save the first three best solutions had been obtained till the moment and make all other searching

agents (including the Omegas) update their positions in accordance to positions of the best searching agents [13, 15-19].

Grey wolves finish the hunting and attack the prey when it stops moving. To model the attack, let us reduce the value of \vec{a} . It should be noted that the fluctuation range of \vec{A} is also reduced at \vec{a} . In other words, \vec{A} is a random value in the interval $[-2a, 2a]$, where a is reduced from 2 to 0 during iterations. When random values of \vec{A} are in the interval $[-1, 1]$, the next position of a searching agent can be anywhere between its current position and the prey's position.

In general, grey wolf find solutions in accordance with the Alpha, Beta and Delta wolves positions. They separate in search of the prey and gather to attack it. To represent the separation mathematically, we use random values of \vec{A} greater than 1 and less than -1 and make the searching agent separate with the prey. This is made by scouts and allows the GWO to implement the global search. $|A| > 1$ makes grey wolves to separate with prey expecting to find the prey's trace. One more component of the GWO that helps scouting is \vec{C} containing random values in the interval $[0, 2]$. It provides the prey with random weight coefficients for the purpose of stochastic enhancement ($C > 1$) or weakening ($C < 1$) of the prey's impact on the distance determination [4,7,19].

The search process starts with generation of random populations if grey wolves (alternatives of solutions) in the GWO. During iterations, Alpha, Beta and Delta wolves estimate the possible position of the prey. Each candidate updates its distance to the prey. Parameter a is reduced from 2 to 0 to emphasize the scouting and the attack respectively. Candidates separate from the prey when $|A| > 1$ and gathering in the direction of the prey when $|A| < 1$. The algorithm stops after the end-point criterion is obtained.

The GWO algorithm includes the following steps:

- 1) To generate the initial population of chromosomes X_i ($i = 1, 2, \dots, n$).
- 2) To initialize the parameter a . At this stage we determine the quantity of mutations providing the effective scouting and attack.
- 3) To calculate the fitness function of each searching agent.
- 4) $X_\alpha, X_\beta, X_\delta$ – are the most promising agents at the moment. Assuming that $X_\alpha, X_\beta, X_\delta$ find and encircle the prey in the most effective way. For this purpose, we range all the chromosomes in accordance with the fitness function ascending. $X_\alpha, X_\beta, X_\delta$ are the first three chromosomes.
- 5) To set the quantity of iterations in terms of the input data. When this parameter is obtained, we fix the value of X_α .
- 6) To update the positions of the wolves X_{t+1} around the prey in relation to $X_\alpha, X_\beta, X_\delta$. For this purpose, we implement the ordering crossover in accordance with the following rule: $X_\alpha * X_\beta \rightarrow X^*$; $X^* * X_\delta \rightarrow X_{t+1}$. Thus, X_{t+1} has the properties of $X_\alpha, X_\beta, X_\delta$.
- 7) To update the parameter a , which is reduced during the iterations narrowing the search area. In other words, the parameter a simulates the attack.

8) To calculate the fitness function of each searching agent in terms of each iteration for the purpose of the further fixation of the best solution.

9) To update X_α, X_β и X_δ as the most promising solutions.

10) If the 'number of iterations' criterion is obtained, we fix X_α as the best solution.

IV EXPERIMENTS

Experimental research was carried out with the different graphs. One of the main problems in terms of bioinspired algorithms is finding the optimal parameters which provide the most effective solutions [20].

To determine time complexity of the GWO algorithm, we carried out a set of experiments and obtained the following results shown in the Fig.2.

The connection between the GWO algorithm time complexity and quantity of elements shown on the diagram affirms the assumed time complexity of the developed GWO algorithm (Fig.2). The GWO algorithm time complexity is of quadratic character and can be represented as $O(\alpha * n^2)$.

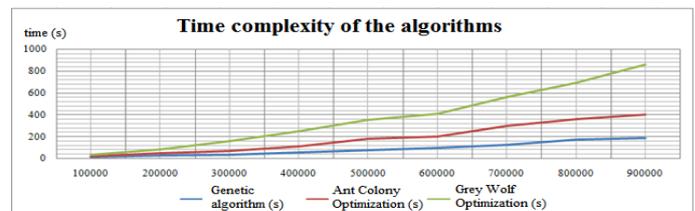


Fig. 2. Dependence of the algorithms time complexity on the quantity of the elements

To estimate the effectiveness of the developed algorithm we carried out the experiments of the solution quality on benchmarks. Let us consider the effectiveness of the algorithm as the quality of solutions obtained with the use of the algorithm. The Fig.3 shows the diagram of the solution quality comparison

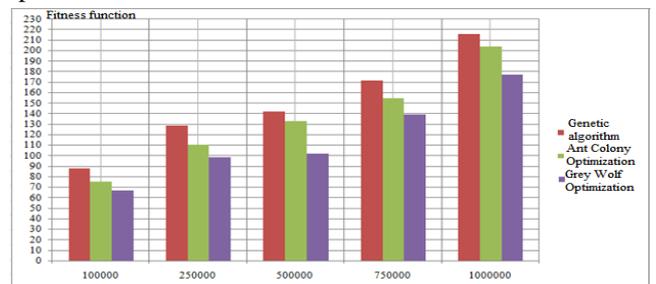


Fig. 3. Diagram of the solution quality comparison

As we can see on the experimental results, the developed GWO algorithm is the most effective one. It is 12% more effective than the genetic algorithm and 8% more effective than the Ant Colony Optimization (ACO) algorithm.

V CONCLUSION

The paper considers the Grey Wolf Optimization (GWO) algorithm. In general, the algorithm is quite effective, but, being a population algorithm, it is not a universal method for solving all the optimization problems.

There are four types of grey wolves: Alpha, Beta, Delta and Omega which are used for modeling of the leadership hierarchy. There are also three main stages of hunting: tracking, encircling and attacking the prey. Which are implemented in the optimization process. Grey wolves belong to the Canidae family and are at the top of the vermin hierarchy, i.e. at the top of the food chain. Grey wolves mostly live in packs consisting of 5-12 species.

The elders are the experienced wolves (Alphas and Betas in the past). Hunters help Alpha and Beta in hunting. Originally, the algorithm is targeted at solving the vector tasks, thus, the authors adapted its main principles for solving the placement task. The initial population is represented as chromosomes, and encoded/decoded in accordance with the following rule: the number of genes in the chromosomes are seats, and the value of genes are numbers of elements placed at the seats.

The paper describes the mathematical representation of the encircling wolves' behavior. To model the hunting behavior, we assume that the Alpha (candidate for the best solution), beta and Delta see the potential position of the prey more clearly. Thus, we save the first three best solutions had been obtained till the moment, and make all other searching agents (including the Omegas) update their positions in accordance to positions of the best searching agent.

Based on the suggested algorithm, the authors developed a software using C++. The main idea of the experimental research is finding a set of the stated task parameters which provide finding quasioptimal solutions in polynomial time. The experiments are carried out on different graphs. Time complexity of the developed GWO algorithm is of quadratic character and can be represented as $O(\alpha \cdot n^2)$. To estimate the effectiveness of the developed algorithm we carried out the experiments on benchmarks of the Ant Colony Optimization (ACO) algorithm and genetic algorithm. The developed GWO algorithm is 12% more effective than the genetic algorithm and 8% more effective than the ACO algorithm.

REFERENCES

- [1] Norenkov I. P. Evolyutsionnyye metody v zadachakh vybora proyektnykh resheniy / I.P. Norenkov, N.M. Arutyunyan // Elektronnoye nauchno-tekhnicheskoye izdaniye
- [2] Karpenko A. P. Sovremennyye algoritmy poiskovoy optimizatsii. Algoritmy, vdokhnovlennyye prirodoy : uchebnoye posobiye / A. P. Karpenko. – Moskva: Izdatel'stvo MGTU im.N. E. Baumana, 2014. – 448 c.
- [3] Akhmedova, SH.A. Ob effektivnosti «staynogo» algoritma optimizatsii. Trudy XLIII Kravoy nauchnoy studencheskoy konferentsii po matematike i komp'yuternym naukam. Krasnoyarsk: SFU, 2010. – S. 9-12.
- [4] Kureychik V.V., Zaporozhets D.YU. Sovremennyye problemy pri razmeshchenii elementov SBIS // Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskoye nauki – 2011. T. 120. № 7. S. 68-73.
- [5] Zaporozhets, D.U., Zaruba, D.V., Kureichik, V.V. Representation of solutions in genetic VLSI placement algorithms – (2014) Proceedings of IEEE East-West Design and Test Symposium, EWDTs 2014.
- [6] Yurevich Zaporozhets, D., Victorovna Zaruba, D., Kureichik, V.V. Hybrid bionic algorithms for solving problems of parametric optimization // (2013) World Applied Sciences Journal, 23 (8), pp. 1032-1036.
- [7] Kuliev, E.V., Dukkardt, A.N., Kureychik, V.V., Legebokov, A.A. Neighborhood research approach in swarm intelligence for solving the optimization problems // (2014) Proceedings of IEEE East-West Design and Test Symposium, EWDTs 2014
- [8] Kureychik V.V., Zaporozhets D.YU. Royevoy algoritm v zadachakh optimizatsii // Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskoye nauki – 2010. T. 108. № 7. S. 28-32.
- [9] Bova, V.V., Lezhebokov, A.A., Gladkov, L.A. Problem-oriented algorithms of solutions search based on the methods of swarm intelligence // (2013) World Applied Sciences Journal, 27 (9), pp. 1201-1205.
- [10] Zaruba, D., Zaporozhets, D., Kureichik, V. VLSI placement problem based on ant colony optimization algorithm // (2016) Advances in Intelligent Systems and Computing, 464, pp. 127-133.
- [11] Kureichik, V., Kureichik, V., Bova, V. Placement of VLSI fragments based on a multilayered approach // (2016) Advances in Intelligent Systems and Computing, 464, pp. 181-190.
- [12] Kureichik, V.V., Zaruba, D.V. The bioinspired algorithm of electronic computing equipment schemes elements placement // (2015) Advances in Intelligent Systems and Computing, 347, pp. 51-58.
- [13] Zaporozhets, D., Zaruba, D.V., Kureichik, V.V. Hierarchical approach for VLSI components placement // (2015) Advances in Intelligent Systems and Computing, 347, pp. 79-87.
- [14] Zaporozhets, D.U., Zaruba, D.V., Kureichik, V.V. Representation of solutions in genetic VLSI placement algorithms // (2014) Proceedings of IEEE East-West Design and Test Symposium, EWDTs 2014.
- [15] Kuliev E.V., Lezhebokov A.A., Dukkardt A.N. Podkhod k issledovaniyu okrestnostey v roevykh algoritmakh dlya resheniya optimizatsionnykh zadach // Izvestiya YUFU. Tekhnicheskoye nauki. Tematicheskoye vypusk «Intellektual'nyye SAPR». – Taganrog: Izd-vo YUFU, 2014. – № 7 (156). – S. 15-26
- [16] Kureychik V.M. Resheniya zadachi razmeshcheniya na osnove evolyutsionnogo modelirovaniya [Tekst] / V.M. Kureychik, B.K. Lebedev, O.B. Lebedev // Izvestiya RAN. Teoriya i sistemy upravleniya. 2007. -№4. - S. 78-91.
- [17] Kureychik V. V., Kureychik V.I. Bionicheskiy poisk pri proyektirovani i upravlenii // Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskoye nauki – 2012. T. 136. № 11 (136). S. 178-183.
- [18] Kureychik V. V., Kureychik V. M., Rodzin S. I. Teoriya evolyutsionnykh vychisleniy. – M.: FIZMATLIT, 2012.
- [19] Kuliev E.V., Lezhebokov A.A. O gibridnom algoritme razmeshcheniya komponentov SBIS // Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskoye nauki. 2012. T. 136. № 11 (136). S. 188-192.
- [20] Kuliev E.V., Lezhebokov A.A. Issledovaniye kharakteristik gibridnogo algoritma razmeshcheniya // Izvestiya YUFU. Tekhnicheskoye nauki. 2013. - № 3. – s. 255- 261.

Intelligent Flow Meter on Acoustic Multivibrator

Zh. A. Sukhinets
Department of
telecommunication systems
Ufa State Aviation Technical
University
Ufa, Russia
sukhinets@mail.ru

A. I. Gulin
Department of automation of
technological processes and
production
Ufa State Petroleum
Technological University
Ufa, Russia
gulin1940@gmail.com

O.I. Bureneva
Department of computer
engineering
Saint-Petersburg State
Electrotechnical University
"LETI"
Saint-Petersburg, Russia
oibur@mail.ru

N. N. Prokopenko
Department Information systems
and radio engineering
Don State Technical University
Rostov-on-Don, Russia
prokopenko@sssu.ru

O. O. Valiamova
Department of automation
of technological processes
and production
Ufa State Petroleum
Technological University
Ufa, Russia
oopoz@mail.ru

Abstract— The article deals with the construction of intelligent flow meters, in which primary detector, sensors, analog-to-digital converters, and microprocessors for solving the problem of distribution of functions between the elements of measurement, conversion and calculation systems are integrated. The flow meter contains a jet generator, which is an acoustic multivibrator. Jet generator converts the flow rate into acoustic vibrations, which are converted into a polyharmonic electrical by piezoelectric transducer. An adaptive line enhancer contains an electronically controlled filter on the varicap and allocates an informative first harmonic and functionally processes it in accordance with the analytical expression of the flow rate from the frequency. The microprocessor processes all information from the adaptive line enhancer and measurement channels from the inert gas concentration sensors, which are impedance biological sensors, and from the integral temperature sensor. The measurement results are displayed on the flow meter display and transmitted via digital communication channels to the dispatcher console.

Keywords— gas flow, jet generator, frequency, concentration, temperature, microcontroller

I. INTRODUCTION

The creation of intelligent gas flow sensors (IS) in which analog-to-digital converters and microprocessors [1, 2], which solve the problem of distribution of functions between the elements of control systems are integrated, makes it possible to process information in the sensor according to a certain algorithm. In addition, there is a possibility of entering corrections from the respective sensors for temperature compensation, the concentration [3-6] of nitrogen, hydrogen sulfide, etc. also, the linearization characteristics to accept commands and transmit the measured values in digital form.

The result is a real metrological characteristics of IS are substantially higher characteristics of the sensors of the

traditional performance. All this contributes to the development and implementation of advanced methods of gas flow measurement, which require significant computational processing, implemented in the sensor microprocessor.

Based on these methods, sensors allow:

- to eliminate moving elements, which increases the reliability of its operation and simplifies maintenance;
- to simplify the requirements for the design of the object of measurement, which extends the use of sensors in different places of objects and reduces the cost of their installation;
- to use new methods and principles of measurement that require quite complex computational processing of sensor output signals, but having a number of advantages in accuracy, stability of readings, ease of installation and maintenance of the sensor during its operation;
- to amendments on temperature, humidity, concentrations of related gases, etc., using the computing device IS;
- to process the information and give the current values of the measured value in the specified units.

II. STATEMENT OF THE RESEARCH PROBLEM

The objective of the study is to develop a method for continuous measurement of gas flow in the pipeline with automatic correction of temperature, humidity, nitrogen concentration, hydrogen sulfide, etc. from built-in nanosensors.

The study has been carried out at the expense of the grant from the Russian Science Foundation (Project No. 16-19-00122-P).

Implementation of the method (Fig. 1) is connected with the solution of several tasks: development of the primary and secondary converters, the functional converter and an information processing unit.

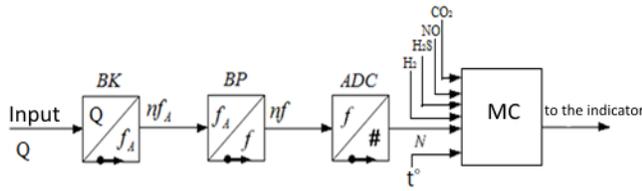


Fig. 1. Block diagram of gas flow measurement

The device consists of a primary converter BK (jet multivibrator), installed in the gas stream, the flow rate Q which is measured, it is excited acoustic oscillations with a frequency f_A , converted by a secondary piezoelectric transducer BP in the corresponding electrical oscillations, representing a series of Fourier frequencies nf , which in turn come to the functional frequency-to-number converter ADC.

The functional converter [7, 8] is an adaptive selector that allocates the first harmonic of the polyharmonic signal, simultaneously functionally converting the frequency into a code corresponding to the measured temperature. The microcontroller (MC), which receives information about the temperature and concentration of the accompanying gases through the appropriate measuring channels, calculates the true gas flow rate, for example methane.

III. SELECTION THE TYPE OF THE JET GENERATOR

The use of a jet generator (JG) – an acoustic multivibrator as a primary converter has a number of advantages, such as the output frequency signal, the lack of moving parts, the relative simplicity of the design, insensitivity to pneumatic shocks.

By type of feedback (FB), causing self-oscillating mode [9], JG are: with the internal FB (Fig. 2, a); with external FB (Fig. 2, b); with two resonance chambers (Fig. 2, c), and with four resonance chambers (Fig. 2, d).

On the example of JG with external FB (Fig. 2, b), consider the process of self-oscillations. The stream flowing from the supply nozzle 1 initially (due to the natural asymmetry of the design) deviates to one of the side walls of the working chamber 2, while increasing the inlet pressure, for example, in the channel FB 3. The interaction of the main flow flowing from the supply nozzle 1 with the control flow flowing from the FB 3 channel leads to the displacement of the main flow from this wall and its adjunction to the opposite wall, and so on. Thus, in the FB 3, 4 channels and in the output channels 5, 6, pressure oscillations are created with a frequency determined by the time delays in the development of processes in the working chamber 2 and the FB 3 and 4 channels.

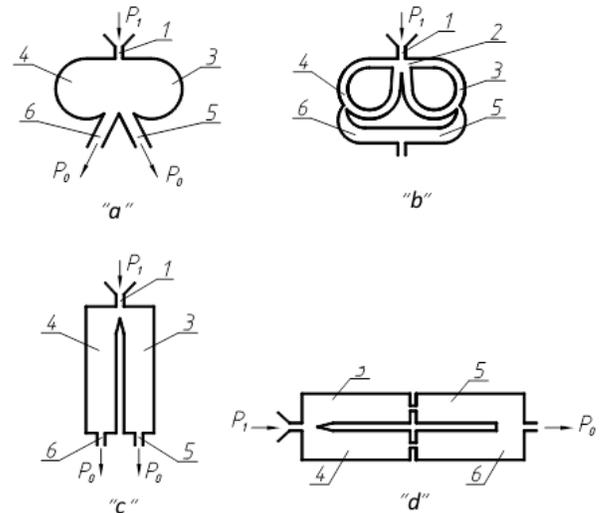


Fig. 2. Schemes of jet generators

Experimental studies of various JG have shown [10] that the frequency of oscillations in them, under other constant conditions, is proportional to the velocity of propagation of disturbances in the gas filling the working cavity and the FB channels, that is:

$$f = \frac{\sqrt{kRT}}{\lambda}, \quad (1)$$

where: T – the absolute braking temperature of the gas at the inlet to the JG;

R – gas constant (for air $R= 287.14 \text{ j}/(\text{kg} \cdot \text{K})$);

k – adiabatic index (for air $k = 1.4$);

λ – the wavelength of the acoustic oscillations JG, equal to the total length of the feedback channels 3 and 4 sensor in Fig. 2, b.

The principle of creating an acoustic multivibrator of oscillations with a frequency proportional to the gas flow rate is used in jet flow meters:

$$f = S_h \frac{Q}{V}, \quad (2)$$

where S_h – Strouhal number;

Q – measured flow rate;

V – the volume of the resonant chamber and FB channels.

In accordance with [11], the frequency of flow oscillations is proportional to the flow rate through the JG nozzle:

$$Q = \mu S \sqrt{2\Delta p / \rho}, \quad (3)$$

where μ – the flow coefficient;

S – the cross-sectional area of the nozzle;

Δp – differential pressure;

ρ – density of the measured medium.

To form a steady-state flow, characterized by a stable regime with an acoustic Reynolds number Re_a in the linear range, straight sections of a certain length are required, excluding the presence of local flow disturbances:

$$\text{Re}_a = \frac{\rho c_0 A}{2\pi fb}, \quad (4)$$

where c_0 – the speed of sound in a medium;
 A – the amplitude of vibrational velocity;
 b – the dissipation parameter, i.e. the transition of the energy of ordered processes into the energy of disordered processes, ultimately – heat.

IV. THE CHOICE OF MATERIAL AND THE OSCILLATING SYSTEM OF THE PIEZOELECTRIC TRANSFORMER

The piezoelectric transformer (PET) has a number of requirements for sensitivity, bandwidth and finding the eigen resonance frequency outside the operating range. The latter condition is necessary to prevent amplification of the second and third harmonics suppressing the informative signal.

The software product ANSYS [12], which has wide possibilities of virtual design and analysis of complex structures, is effective in the development of the PET. Modal analysis allows to analyze the design. Harmonic analysis allows not only to clarify the values of resonant frequencies, but also to determine the operating frequency ranges and sensitivity.

To match the wave resistances of the gas medium and piezoceramics, a thin-walled semi-passive bimorph with a resonance frequency of about 40 kHz was selected from the zirconate titanate, which has a temperature range of up to 200 °C, the required capacitance of 0.1 nF, and an Electromechanical coupling coefficient of 0.4, which will achieve the required frequency band at a sensitivity of about 50 $\mu\text{V}/\text{Pa}$.

V. SELECTION OF ADAPTIVE SELECTOR

The polyharmonic signal after the PET, the nonlinear dependence of the output frequency of JG on the physical parameters do not allow the use of standard frequency schemes in gas flow measurement systems. In this regard, it is recommended to use as an adaptive frequency selector nominal values [4], with the introduction (Fig. 3) additional measuring channels for temperature, etc. The adaptive selector operates as follows. The input polyharmonic f_x signal is fed to an electronically controlled phase shifter (ECPS), tunable by a sawtooth-voltage generator (SG), and connected to the first phase comparator (PC) input via an amplifier. The second input of the PC receives a signal directly from the output of the PET. ECPS carries out frequency tuning to equal phases with the frequency of the first harmonic f_1 , coming directly to the second input of the comparator, under the control of the SG, launched by a single vibrator (SV). PC in case of equality of the phases through the key records the voltage sweeps SG and produces a signal to close the circuit matches "AND", stopping the pulse counter. The meter data is fed to the microcontroller (MC) for functional information processing in accordance with the analytical ratio of flow rate to frequency.

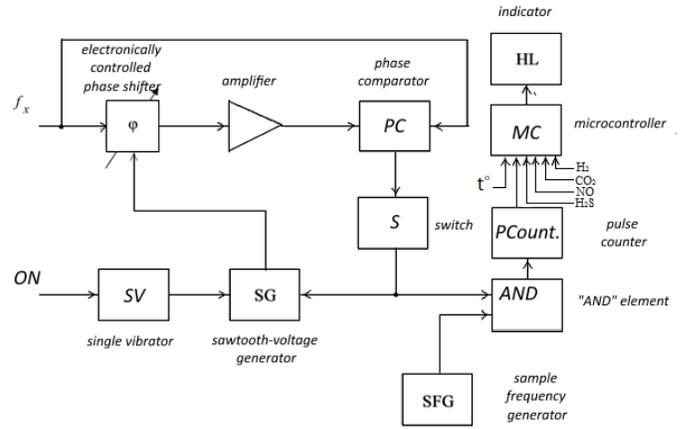


Fig. 3. Adaptive selector

The use of phase-locked frequency (PLF) for ECPS, consisting of RC-links, where the role of tunable capacitances perform varicaps, increases accuracy, since there is no methodological error at the time of frequency measurement [13], and speed because varicaps, virtually non-inertial elements to the submillimeter range, and analog signal processing without intermediate transformations, greatly simplifies the scheme, which increases the reliability of the system. Recommendations for the calculation of ECPS are set out in [14].

VI. AUTOMATIC CORRECTION OF THE CONCENTRATION OF INERT GASES

The use of [14-20] biological sensors (BS) in devices implementing the method of electrochemical impedance spectroscopy (EIS) allows detecting and measuring the concentration of various gases. The data obtained from the EIS devices are processed in the MC. The EIS device together with the MC form a single specialized electrochemical impedance spectroscopy (SEIS) system, which has a large number of built-in databases for various types of gases, for example, H_2 , CO_2 , NO , H_2S and can be implemented as a system on a chip with internal integration of functional blocks.

In [21-26], the general structure and circuitry of the main functional units of the SETS based on the amplitude-phase method, which allows to improve the basic metrological parameters and simplify the analog interface, are investigated.

The scheme of BS parameters measurement in the high-frequency EIS device using peak and phase detectors is shown in Fig. 4.

The voltage at the output of the phase detector is proportional to the phase difference between the signal from the amplifier output and the polling signal.

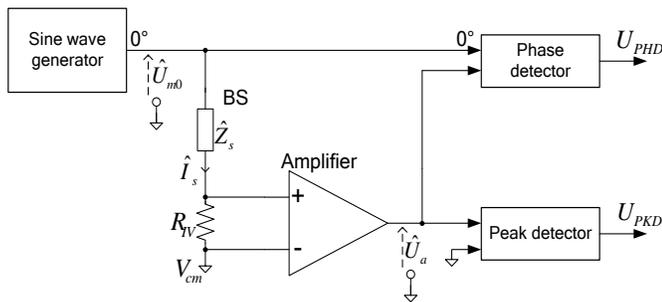


Fig. 4. Functional scheme of BS parameters measurement based on peak and phase detectors

Therefore, the required argument of the complex impedance of the DB can be found as:

$$\arg Z_s = -\varphi = -\frac{U_{PHD}}{K_{PHD}}, \quad (5)$$

where U_{PHD} – the output voltage of the phase detector;
 K_{PHD} – some given coefficient of proportionality between the input phase difference and the output voltage.

The peak detector determines the amplitude of the amplifier signal U_a , which in turn characterizes the amplitude of the BS current I_s . Using this, we can express the desired module of the complex impedance of the BS:

$$|Z_s| = \frac{U_{m0}}{I_s} = \frac{AU_{m0}}{R_{IV}U_{PKD}}, \quad (6)$$

where A – the gain;
 I_s – module the output of the integrated current BS;
 R_{IV} – constant resistance for converting current-to-voltage;
 U_{m0} – the voltage amplitude of the polling signal.

The output values U_{PHD} and U_{PKD} change slowly due to the low rate of chemical processes inside the BS. This allows the use of relatively low-frequency ADC and digital blocks to calculate the phase and impedance module.

VII. AUTOMATIC CORRECTION OF THE GAS TEMPERATURE

New integrated temperature sensors allow to change the approach to the design of information systems due to the high characteristics [27], ease of use in microprocessor measurement devices and automation, including gas flow meters. Example of new advanced sensors it can serve as LMT970 Texas Instruments. It has a temperature range of $-55\text{ }^{\circ}\text{C}$ to $+150\text{ }^{\circ}\text{C}$ and an accuracy of $\pm 0.05\text{ }^{\circ}\text{C}$. Its analog output voltage is usually digitized by the ADC of the accompanying microcontroller.

CONCLUSION

1. Common ultrasonic flowmeters of "Teplopribor" OOO (Ryazan), "DIMET" ZAO (Tyumen), "Irvis" OOO (Kazan), Controtron and GE Panametric (USA), SICK/Maihak GmbH (Germany) and others contain from two to eight piezoelectric emitters to improve the accuracy of flow measurement, which complicates the information

processing scheme, increases power consumption and reduces reliability.

2. Information on the concentration of inert gases is obtained periodically from a separate expensive chromatograph.

3. The proposed flow meter on the acoustic multivibrator has much smaller dimensions, does not contain piezoelectric emitters, and due to the built-in integrated sensors of inert gas concentration allows to continuously subtract their volume from the total flow rate.

REFERENCES

- [1] Ngo Ch., van de Voorde M. Nanotechnology in a Nutshell. Atlantis Press, 2014, 502 p. ISBN-13: 978-94-6239-011-9.
- [2] Shkola dlya elektrika. Available at: <http://electricalschool.info/automation/1829-intellektualnye-datchiki-ikh.html> (accessed 7 March 2018).
- [3] MIFI: MDP sensors D-1. Available at: <https://sensor.mephi.ru/sensors.htm> (accessed 7 March 2018).
- [4] Yan, M., Nan, X., Liru, H., Jinghua, T., Weimin, C. Development on Intelligent Small-Flow Target Flow Meter. Computer Science-Technology and Applications, International Forum on, ChongQing, China, 2009, pp. 315-317. doi: 10.1109/IFCSTA.2009.317
- [5] High-Precision Flow Measurement for an Ultrasonic Transit Time Flowmeter. 2010 International Conference on Intelligent System Design and Engineering Applications, Changsha, Hunan China, 2010, pp. 823-826. doi: 10.1109/ISDEA.2010.138
- [6] Stanley, M., Gervais-Ducouret, S., Adams, J. Intelligent sensor hub benefits for wireless sensor networks. 2012 IEEE Sensors Applications Symposium Proceedings, Brescia, Italy, 2012, pp. 1-6. doi: 10.1109/SAS.2012.6166299
- [7] Sukhinets Zh.A., Gulin A.I. Functional nominal frequency values of sinusoidal signals for frequency detectors. *Pribory i sistemy. Upravleniye. Kontrol. Diagnostika* [Devices and systems. Management, Control, Diagnostics]. Moscow, 2012, no. 9 – p. 33-37.
- [8] Sukhinets Zh.A., Gulin A.I. *Sposob izmereniya nominalnoy chastoty sinusoidalnykh signalov i ustroystvo dlya ego realizatsii* [The way to measure nominal frequency of sinusoidal signals and device for its implementation] Patent RF, no. 2503019, 2013.
- [9] Zalmanson L. A. *Teoriya elementov pnevmoniki* [Theory of elements of pneumonics]. Moscow, Nauka Publ., 1969, 508 p.
- [10] Zolotarevsky S. A. On the applicability of various flow measurement methods for commercial gas metering. *Energoanaliz i energoeffektivnost'*, 2006, no. 2(15).
- [11] Ivanushkin I. Yu. On the applicability of the jet method in the measurement of gas flow. *Sbornik statey «Kommercheskiy uchet prirodnogo gaza. Novoye gazoizmeritelnoye oborudovaniye i sistemy»* [Collection of articles "Commercial accounting of natural gas. New gas measuring equipment and systems"], 2011.
- [12] Mitko V.N., Kramarov Yu.A., Panich A.A. *Matematicheskoye modelirovaniye fizicheskikh protsessov v pyezoelektricheskom priborostroyenii: Monografiya* [Mathematical modeling of physical processes in piezoelectric instrument engineering: Monograph]. Rostov-on-the-Don, YuFU Publ., 2009, 240 p.
- [13] Sukhinets Zh. A., Gulin A. I. Modeling of converters and devices with distributed RC-parameters with the required accuracy in the specified frequency range. *2nd international conference on industrial design, application and manufacture (ICIEAM)*, 2016, p. 1-6, doi: 10.1109 / ICIEAM. 2016. 7911677.
- [14] Li L-D, Zhao H-T, Chen Z-B, Mu X-J, Guo L. Aptamer biosensor for label-free impedance spectroscopy detection of thrombin based on gold nanoparticles. *Sensors and Actuators: B Chemical* 2011, pp. 189-194.
- [15] Kim B K, Li J, Im J-E, Ahn K-S, Park T S, Cho S I, Kim YR, Lee W-Y. Impedometric estrogen biosensor based on estrogen receptor alpha-immobilized gold electrode. *Journal of Electroanalytical Chemistry* 2012; 671, pp. 106- 111.
- [16] Lin J., Wang R., Jiao P. et al. An impedance immunosensor based on low-cost microelectrodes and specific monoclonal antibodies for rapid

- detection of avian influenza virus H5N1 in chicken swabs. *Biosens Bioelectron*, 67, 2015, May 19, pp. 546-52.
- [17] Ensafi A.A., Amini M., Rezaei B., Talebi M. A novel diagnostic biosensor for distinguishing immunoglobulin mutated and unmutated types of chronic lymphocytic leukemia. *Biosensors and Bioelectronics* 2016, 77, pp. 409-415.
- [18] Pichetsurnthorn P, Vattipalli K, Prasad S. Nanoporous impedimetric biosensor for detection of trace atrazine from water samples. *Biosensors and Bioelectronics* 2012; 32, pp. 155-162.
- [19] Manickam A., Chevalier A., McDermott M., Ellington A.D., and Hassibi A. A CMOS electrochemical impedance spectroscopy biosensor array for label-free biomolecular detection. *Proc. IEEE Int. Solid-State Circuits Conf.*, 2010, pp. 130–131.
- [20] Prada J., Vega-Castillo P., Krautschneider W. Design of a Wide Tuning-Range CMOS 130-nm Quadrature VCO for Cell Impedance Spectroscopy. *6th IEEE Germany Student Conference Proceedings*, Hamburg, 2015, p. 7-12.
- [21] Jafari H., Soleymani L., and Genov R. 16-Channel CMOS Impedance Spectroscopy DNA Analyzer With Dual-Slope Multiplying ADCs. *IEEE Transactions on Biomedical Circuits and Systems*, 2012, Vol. 6, No. 5, p. 468-478.
- [22] Yang A., Jadhav S.R., Worden R.M., and Mason A.J. Compact low-power impedance-to-digital converter for sensor array microsystems. *IEEE J. Solid-State Circuits*, 2009, Vol. 44, No. 10, p. 2844-2855.
- [23] Helmy A.A. and Entesari K. A 1 – to – 8 GHz miniaturized dielectric spectroscopy system for chemical sensing, *IEEE MTT-S int. Microw. Symp.*, Jun. 2012, p. 493-495.
- [24] Samoilov L.K., E.A. Zhebrun E.A., Titov E.A. Analog Interface Microcircuitry for Electrochemical Impedance Spectroscopy Systems, *VII All-Russia Science&Technology Conference MES-2016. «Problems of Advanced Micro- and Nanoelectronic Systems Development»*, 2017, Part III, Moscow, IPPM RAS, pp. 17-22.
- [25] Samoilov L.K., E.A. Zhebrun E.A., Prokopenko N.N., Budyakov P.S. Research of peak detector limiting characteristics for analog interface in impedance spectroscopy systems. *2017 IEEE International Conference on Electronics, Circuits and Systems, ICECS 2017*, Batumi, Georgia, December 5-8, 2017, pp. 423-426.
- [26] *Gazovyye datchiki i sensory*. Available at: <http://www.gassensor.ru/ru/events> (accessed 15 March 2018).
- [27] *ООО NPP «ELEMER»*. Available at: <https://elemerufa.ru/production/datchiki-temperaturyi/preobrazovatelyi-temperaturyi-i-vlazhnosti/> (accessed 19 March 2018).

Technique to Simulate Oscillator Circuits with the Degradation Models

Mark M. Gourary
CAD department
IPPM RAS
Moscow, Russian
Federation
gourary@yandex.ru

Sergey G. Rusakov
CAD department
IPPM RAS
Moscow, Russian
Federation
rusakov@ippm.ru

Sergey L. Ulyanov
CAD department
IPPM RAS
Moscow, Russian
Federation
ulyas@ippm.ru

Michael M. Zharov
CAD department
IPPM RAS
Moscow, Russian
Federation
zarov@ippm.ru

Abstract—The approach to predict the effect of negative electro-temperature instability (NBTI-effect) in analog VLSI using circuit simulators is proposed. This approach is directed to simulate circuit behavior taking into account the growth of threshold voltage under the long-term effect of negative voltage on the gate of p-MOSFET transistors. A new computational procedure based on the harmonic balance method is developed. In this procedure NBTI models are included as nonlinear elements into the model of the analyzed circuit to perform numerical analysis of the NBTI-effect. In contrast to the known methods, the new approach does not require preliminary estimation of the NBTI model parameters averaged over the signal period. It also provides reducing the computational costs due to selecting the integration method and the integration step. The approach provides oscillator circuit simulation with the NBTI effect. The approach is applicable to multi-frequency circuits.

Keywords—analog integrated circuits, circuit simulation, NBTI effect

I. INTRODUCTION

The effect of instability – so-called NBTI-effect (Negative Bias Temperature Instability) is one of the main factors affecting the reliability of the operation of nanometer CMOS integrated circuits (IC). This effect manifests itself at high temperatures and prolonged exposure of the negative voltage to the gate of the p-MOS transistor. NBTI-effect causes an increase in the threshold voltage, which affects the characteristics of the circuit and can lead to violations of the circuit during long-term operation.

The degradation of the threshold voltage of the p-channel transistor is important effect in submicron MOSFETs. Scaling problems of MOS transistors in modern submicron IC design are closely connected with increasing the internal electric fields. So the thickness of the gate oxide of IC MOS devices has decreased from tens of nanometers to below 2 nanometers for 0.13 and 0.09 μm technologies [1] that led to the growth of the internal electric fields. In this case the negative bias temperature instability (NBTI) of p-MOSFET has been identified as a critical limiting factor that ultimately determines the lifetime of the devices. Therefore VLSI developers should predict this effect in the early stages of design to ensure the correct functioning of the circuit within a given time period.

To successfully predict the results of undesirable aging processes and estimate the time intervals of degradation, it is necessary to use the device models that provide sufficient accuracy of the NBTI-effect description, as well as to develop specialized numerical procedures for circuit simulators.

This article discusses the algorithmic aspects of the development of NBTI-effect impact assessment in circuit simulators.

The most part of the papers on this topic is focused on the estimation of NBTI-effect in digital VLSI (for example, [2-5]). Such computational methods do not apply directly to analog circuits. The limitations are due to the fact that in digital VLSI the static levels are selected as stress voltages on the gates. In analog circuits the stress voltages change over time.

Currently, various models of devices based on the description of physical processes in the oxide causing the NBTI-effect are used [4, 5]. These processes determine the concentration of traps in silicon oxide and on the boundary of silicon-silicon oxide. The time dependence of the concentration is represented as a power function of time with a coefficient depending on the electric field strength in the oxide and temperature [6]. Another so-called diffusion model is presented in [7]. Both models lead to similar characteristics of threshold voltage degradation.

The time dependences of the threshold voltages obtained on the basis of such models correspond to the constant gate voltages of p-MOS devices. However in most cases the gate voltages are not constant and have time-varying character.

The following types of variation are possible:

- 1) fast (high-frequency) variations defined by the device operational mode,
- 2) slow variations defined by DC shift due to the NBTI effect.

Fast variations are usually simulated by introducing an equivalent DC voltage depending on the average gate voltage during the waveform period [8]. One of the aims of this paper is to develop an approach to evaluate the corresponding dependence.

Slow variations can be simply evaluated for externally excited blocks with gate waveform independent on the gate

properties. For oscillator circuits the specified time aging interval is divided into subintervals with constant gate properties [9-11]. In this case the choice of the step size is performed using heuristic procedure.

This paper is directed to avoid heuristic procedure and provide the full automation of analysis using standard circuit simulators.

II. MAIN MODELS FOR NBTI ANALYSIS

The typical time dependence of threshold voltage shift for gate oxide thickness 1.3 nm is given in Fig. 1 [1]. In this figure the dependence of threshold voltage deviation is given in log-log plot. The time dependence exhibits a power law. The deviation of current-voltage characteristic $I_{DS}-V_{DS}$ due to the threshold voltage growth [8] is illustrated in Fig. 2.

The following power dependence is usually applied for numerical evaluation of the threshold voltage (V_{th}) growth under the constant gate-source voltage [12]:

$$\Delta V_{th} = A t_s^n, \quad (1)$$

where ΔV_{th} is the deviation of the threshold voltage during the aging time t_s , n is an exponent closing to theoretical value $n=1/6$, factor A is defined as

$$A = C_0 \exp(\alpha E_{ox}) \exp\left(\frac{E_a}{kT^\circ}\right) = C \exp(\alpha V_{dd} / d_{ox}). \quad (2)$$

Here $E_{ox} = V_{gs} / d_{ox}$ is the field strength in the oxide with the thickness d_{ox} , factors C and α depend on the parameters of technology, $C = C_0 \exp\left(\frac{E_a}{kT^\circ}\right)$, E_a is the activation energy, T° is the temperature.

In digital IC the gate voltage can accept two voltage levels – the supply voltage (Vdd) or 0 (“ground”), which do not depend on the device threshold voltage. Thus there are no slow variations of the gate voltage in such circuits. Switching of voltage levels is taken into account by relative part (γ) of the aging time, when the gate is under high voltage. In this case expression (1) is transformed to [12, 13]:

$$\Delta V_{th} = A (\gamma t_s)^n. \quad (3)$$

It is assumed that (3) does not depend on the switching frequency.

The similar approach cannot be applied to analog circuits where gate voltage is not defined by pulsed waveform. Slow variations can be taken into account by the following modification of (1) presented in [8]

$$\Delta V_{th} = \left[\int_0^{t_s} (A(\tau))^{1/n} d\tau \right]^n. \quad (4)$$

Here $A(\tau)$ is obtained taking into account slow variations of the average gate voltage $\bar{V}_{gs}(\tau)$ which is defined in [8] as follows

$$A(\tau) = A(\bar{V}_{gs}(\tau)) = C \exp(\alpha \beta \bar{V}_{gs} / d_{ox}), \quad (5)$$

where

$$\bar{V}_{gs}(\tau) = \frac{1}{T(\tau)} \int_0^{T(\tau)} V_{gs}(\tau+t) dt, \quad (6)$$

$T(\tau)$ is the waveform period after aging time τ , β is an empirical parameter [8-11]. One can easily see that (4)-(6) coincide with (1) for time-independent $A(t)$.

The integration in (4), (6) can be easily performed for blocks with the external excitation when the gate waveform $V_{gs}(\tau+t)$ is known. For the autonomous circuits variations of V_{th} require more complicated technique. The technique for solving (4)-(6) presented in [9] is based on using the multistep algorithm. At each step the steady-state simulation is performed and factor A is determined from (5)-(6). Then the increment of the gate threshold voltage is obtained from (4).

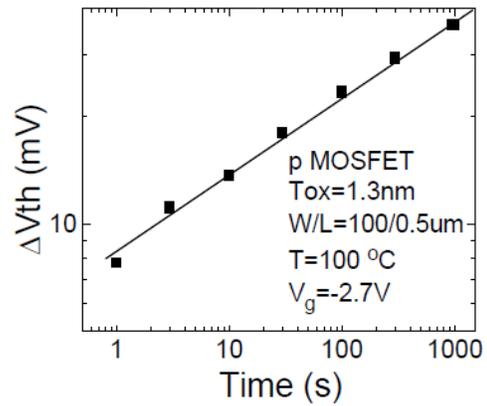


Fig. 1. The time dependence of threshold voltage shift [1].

The following shortcomings of such technique can be pointed out:

- empirical definition of β in (5) complicates the automation of the analysis process,
- the heuristic step size determination in [9] is not based on the rigorous mathematical approach associated with the desired accuracy of the computation,
- the technique is directed only to the periodic mode and can not be applied to the almost periodic multiple-frequency mode.

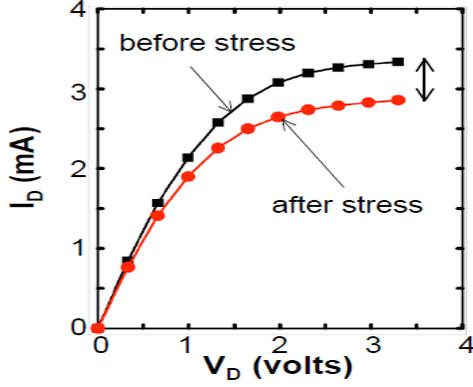


Fig. 2. Illustration of the drain characteristic deviation due to NBTI stressing. [6].

III. A NEW APPROACH TO NBTI SIMULATION

Here we consider a new technique which eliminates above mentioned limitations. The technique is based on the representation of the integral dependence (4) in the form of the differential equation

$$\frac{ds}{d\tau} = (A(\tau))^{1/n}, \Delta V_{th} = s^n, \quad (7)$$

which obtained by the time differentiation of (4).

Assuming the frequency independence of (7) one can conclude that (7) is true at the oscillation period. Taking into account that the threshold voltage is practically constant value for one period, its average on the period can substitute for variable s in (7). Then we obtain

$$\frac{d\bar{s}(\tau)}{d\tau} = \frac{C^{1/n}}{T} \int_0^T \exp(\alpha V_{gs}(t, V_{th}(\tau)) / n d_{ox}) dt, \quad (8)$$

$$V_{th}(\tau) = V_{th}(0) + \bar{s}^n(\tau). \quad (9)$$

Here $V_{gs}(t, V_{th})$ defines the waveform of the gate voltage in the circuit with the threshold voltage V_{th} . Unlike (5) the application of (8), (9) does not require the value of empirical factor β . It can be easily shown that the application of (8), (9) to digital circuits with pulsed gate waveforms leads to (3).

If the circuit contains several p-MOS transistors then (8), (9) are applied to each transistor that leads to ODE system. The system can be solved by any numerical integration method (Euler, trapezoidal, Runge-Kutta, etc.). In the context of representation by ODE (8) the standard step size control algorithms can be applied to provide required accuracy.

The calculation of right hand side (RHS) of ODE (8) requires the numerical evaluation of the integral at $0 \leq t < T$ while circuit simulation. This evaluation can be avoided taking into account that RHS of (8) represents the average of the expression

$$U(t) = C^{1/n} \exp(\alpha V_{gs}(t, V_{th}) / d_{ox}) \quad (10)$$

for interval T . The average can be obtained as a DC component of the output signal of the nonlinear device with transfer function defined by (10). This value can be evaluated by the harmonic balance (HB) method [14] as a zero harmonic of the nonlinear device output.

Since the HB method allows determining constant components of nodal variables in case of multitone excitation the proposed approach can be easily extended to the simulation of multitone steady-state modes of electronic circuits.

IV. NUMERICAL EXAMPLE

Below we consider the application of the proposed technique to the CMOS oscillator circuit presented in Fig. 3. To provide the analysis the nonlinear device (ND) with transfer function (10) is attached to the circuit. The circuit has two p-MOSFETs. Due to the circuit symmetry their threshold voltages vary equally. Hence NBTI analysis of the circuit requires only one ODE (8, 9).

The ODE solving is performed by forward Euler method. At every step the circuit is simulated by the HB method. Numerical results are presented in Fig. 4, 5. Fig. 4 shows the computed time dependences of threshold voltages of p-MOS transistors under NBTI-effect. The resulting decrease of the first harmonic magnitude of the gate voltage can be seen in Fig. 5.

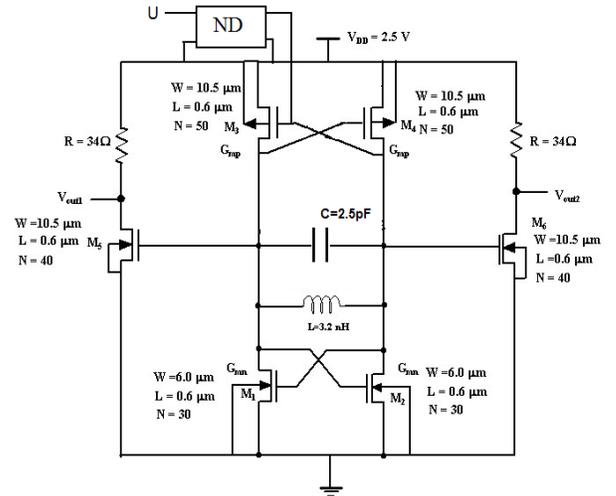


Fig. 3. CMOS oscillator circuit for the analysis of NBTI-effect.

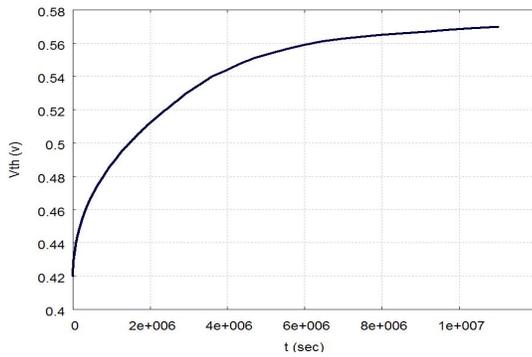


Fig. 4. The computed characteristic of aging of the threshold voltage of p-MOS transistors under NBTI-effect.

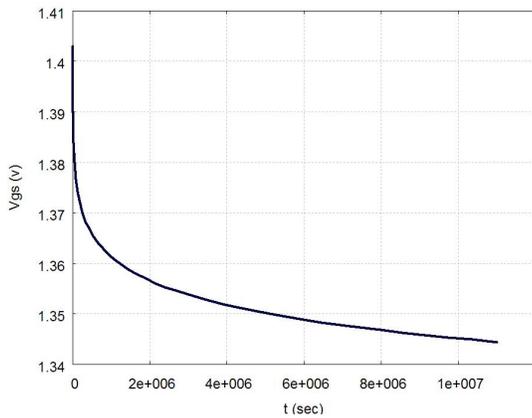


Fig. 5. The computed time dependence of the first harmonic of the gate voltage in p-MOS transistors under NBTI-effect.

V. CONCLUSION

A new approach for the NBTI analysis of oscillators is proposed. The approach produces the following advantages in comparison with known methods:

- the representation of the threshold voltage degradation mechanism in the form of ODE system provides the utilization of standard stepsize control algorithm under numerical integration,

- the evaluation of the RHS of ODE system by the simulation of the oscillator circuit supplemented by the developed nonlinear block allows to eliminate empirical average coefficients,

- the application of the Harmonic Balance method for the oscillator simulation provides the ability to analyze multitone modes.

REFERENCES

- [1] G. Chen, M. F. Li, C. H. Ang, J. Z. Zheng, and D. L. Kwong, "Dynamic NBTI of p-MOS Transistors and its Impact on MOSFET Scaling," *IEEE Electron Device Letters.*, v. 23, Issue: 12, Dec. 2002. pp. 734 – 736.
- [2] X. Li, Quin J., Huang B., Zhang X., J.B. Bernstein, "A New SPICE Reliability Simulation Method for Deep Submicrometer CMOS VLSI Circuits," *IEEE Trans. On Device and Materials Reliability*, v. 6, June 2006, pp. 247-257.
- [3] X. Xuan, A. Chatterjee, A.D.Singh, N.R. Kim and M.T. Chisa, "IC reliability simulator ARET and its application in design-for-reliability," *Proc. Asian Test Symposium*, Xian, China, v. 12, 2003, pp. 18-21.
- [4] A. Stempkovsky, A. Glebov, S. Gavrilov "Calculation of Stress Probability for NBTI-Aware Timing Analysis," *Proc. Conf. ISQED*, 2009, pp. 714-718.
- [5] U. Dutta, M.K. Soni, M. Pattanaik, "A Review of NBTI Degradation and its Impact on the Performance of SRAM", *Int. J. of Modern Education and Computer Science (IJMECS)*, vol.8, No. 6, 2016. pp.57-65. DOI: 10.5815/ijmecs.2016.06.08.
- [6] M. Alam, S. Mahapatra, "A comprehensive model of PMOS NBTI degradation," *Microelectron. Reliab.*, vol. 45(1), 2005, pp. 71–81.
- [7] J.H. Stathis, S. Zafar, "The negative bias temperature instability in MOS devices: A review," *Microelectron. Reliab.*, vol. 46(2-4), 2006, pp. 270-286.
- [8] E. Maricaeu, G. Gielen, "Efficient reliability simulation of analog ICs including variability and time-varying stress," *Proc. DATE Conf.*, April 2009, pp. 1238-1241.
- [9] E. Maricaeu, G. Gielen, "Efficient Variability-Aware NBTI and Hot Carrier Circuit Reliability Analysis," *IEEE Trans. Comp.-Aided Design of Integ. Circ. and Sys.*, vol. 29(12), December 2010, pp. 1884-1893.
- [10] E. Maricaeu and G. Gielen, "Computer-aided analog circuit design for reliability in nanometer CMOS," *IEEE J. Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 1, pp. 50–58, Mar. 2011.
- [11] G. Gielen, E. Maricaeu, P. De Wit, "Analog circuit reliability in sub-32 nanometer CMOS: Analysis and mitigation," *Proc. DATE Conf.*, 2011, pp. 1474-1479.
- [12] K. Ramakrishnan, X. Wu, N. Vijaykrishnan, Y. Xie, "Comparative analysis of NBTI effects on low power and high performance flip-flops," *Proc. of IEEE Int. Conf. on Computer-aided Design*, 2008, pp. 200-205.
- [13] Wenping Wang, Zile Wei, Shengqi Yang, Yu Cao, "An efficient method to identify critical gates under circuit aging," *Proc. of IEEE/ACM Int. Conf. on Computer-aided Design*, 2007, pp. 735-740.
- [14] K.S. Kundert, J. White, A. Sangiovanni-Vincentelli, *Steady-State Methods for Simulating Analog and Microwave Circuits*, Kluwer Academic Publishers, Boston, 1990.

Polynomial Code with Detecting the Symmetric and Asymmetric Errors in the Data Vectors

Ruslan B. Abdullaev,
“Automation and Remote Control on Railways” Department,
PhD student,
Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
ruslan_0507@mail.ru

Valerii V. Sapozhnikov,
“Automation and Remote Control on Railways” Department,
DSc, Professor,
Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
port.at.pgups@gmail.com

Dmitrii V. Efanov,
“Automation, Remote Control and Communication
on Railway Transport” Department,
DSc, Professor,
Russian University of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

Vladimir V. Sapozhnikov,
“Automation and Remote Control on Railways” Department,
DSc, Professor
Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
at.pgups@gmail.com

Abstract—Paper contains results of researches in polynomial codes field at error detecting in data vectors. The authors established previously unknown polynomial codes properties, which consideration is expedient when organizing systems with fault detection. A separable binary codes qualitative classification is given, it is shown that special code classes can be distinguished that are focused on the certain types errors detection arising in the data vectors (unidirectional, symmetric or asymmetric). Examples of codes included in each special codes class, as well as methods for their use in the automation and computing systems construction, are given. The generator polynomial types that allow detecting all symmetric and asymmetric errors in the polynomial code data vectors are established. Some experimental studies result with a combinational circuits benchmarks LGSynth'89 set, confirming the results of theoretical studies are given.

Keywords—self-checking structures; fault detection systems; binary errors detection; polynomial codes, unidirectional errors, symmetrical errors, asymmetric errors; separate type errors detection.

I. INTRODUCTION

The rapid development of intelligent technologies happens against the background of constantly becoming complicated and improved technical components. Today, they have such miniature dimensions that even developers of transistors could not reflect on its last century [1 – 3]. The various stages of technology evolution bring in some difficulties in ensuring the reliable and safe operation on the final systems. The main role for the effective operation of modern systems is played by methods and means of technical diagnosing and status monitoring of blocks and nodes as well as ensuring failsafe work of components [4, 5].

The most important means of ensuring high reliability and efficiency of using modern technology is noise-resistant coding of information at all levels of its implementation. The principles of noise-proof coding are used when entering hardware redundancy into the architecture of managing systems when choosing backup methods for components and, they are applied at protection and processing of control data [6 – 8]. When building a highly reliable device, the selected encoding method determines its characteristics. In some

tasks, error correction properties (for example, when transmitting data over distances) are important, and in some, error detection properties (for example, in hardware implementations of devices to prevent the accumulation of faults).

This paper is devoted to presenting the results of studies on detection errors in data vectors by polynomial codes [9]. Presents classes of polynomial codes with detection errors of certain types in the data vectors, which is fundamental when choosing a coding method at the design stage of a highly reliable device.

II. CODES WITH DETECTION ERRORS OF VARIOUS TYPES

Often, the properties of detecting errors of various types are used in the implementation of technical diagnostic tools and the construction of systems with controllable architectures [10, 11]. Classical constant-weight codes and Berger's codes that have the property of detection of any unidirectional manifestations of distortions are widely used [12]. Unidirectional is such errors in code words (or separately data vectors), which occur in the presence of distortions of only zero or only one bit. All other errors are non-unidirectional. In the set of non-unidirectional errors, symmetric and asymmetric errors are distinguished. Symmetric errors occur with the same number of distortions of 0 and 1 digits, and asymmetric errors occur with a different number of such distortions. In [13], it was shown that with an increase in the magnitude of the error rate, the proportion of unidirectional errors in their total number gradually decreases, and asymmetric errors increase. The proportion of symmetric errors with increasing multiplicity d decreases, but not as rapidly as unidirectional errors. These features can be considered in the development of reliable systems with fault detection.

The selection of symmetric errors from a set of non-unidirectional errors is since all such errors will not be detected by classical Berger codes and some of them by constant-weight codes. The remaining types of errors will be detected by these codes. In [14], it was shown that for any symmetric errors to be detected, any redundant code must have high redundancy, which is not commensurate with the

redundancy of the same Berger codes. Detection of unidirectional errors requires much smaller redundancy (their share in total number much less). For example, Berger codes have $k = \lceil \log_2(m+1) \rceil$ check bits (here, m – is the number of data bits). Constant-weight codes and Berger codes form a class of so-called $UAED(m,k)$ -codes (unidirectional and asymmetrical error-detection codes). They can be used without special restrictions, for example, when constructing external check circuits for logic devices with such structures, at the outputs of which only unidirectional and asymmetric manifestations of faults are possible. Algorithms for transforming automation circuits into circuits of this type are known: both separately with permissible monotonic distortions [15 – 18], and with permissible and unidirectional and asymmetric distortions at the outputs [19, 20].

When reducing redundancy, for example, with respect to the classical Berger codes, the detection properties of any unidirectional and asymmetric errors are lost. The codes obtained in this way no longer belong to the class $UAED(m,k)$ -codes. However, for many methods of constructing codes, the limiting multiplicities of undetectable unidirectional (d_v) and asymmetric errors (d_a) can be distinguished. We denote this class of codes as d_v, d_a - $UAED(m,k)$ -codes, where d_v and d_a are those minimum values of multiplicities for which undetectable unidirectional and asymmetric errors occur. Examples of d_v, d_a - $UAED(m,k)$ -codes are the classic and modified Bose-Lin codes [21, 22]. For such codes, the values of d_v and d_a are fixed, and the codes themselves are $M, (M+2)$ - $UAED(m,k)$ -codes, where M – is the value of the module for calculating deductions when generating code words.

However, for some special cases, it is possible to build codes that detect any symmetric errors, so-called $SED(m, k)$ -codes (or d_s - $SED(m,k)$ -codes, where d_s is the minimum multiplicity of undetectable symmetric errors) [14]. The application of such codes can be like the $UAED$ application (m,k)-codes, but with check of devices on other property – the symmetry of distortions.

Taking into account the peculiarities of application codes when building systems with detection of faults, one can distinguish classes of codes with detection of any unidirectional and with detection of any asymmetric errors – $UED(m,k)$ and $AED(m,k)$ -codes (including d_v - $UED(m,k)$ and d_a - $AED(m,k)$ -codes). The classification of codes by the properties of detecting various errors is shown in Fig. 1.

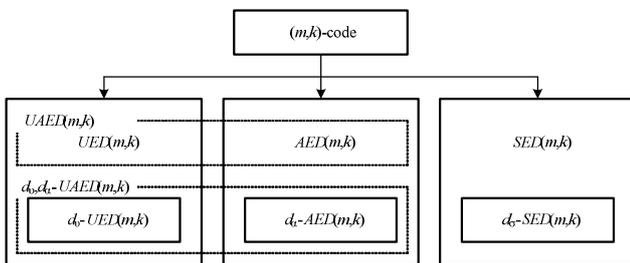


Fig. 1. Classification codes.

III. POLYNOMIAL CODES WITH SYMMETRIC AND ASYMMETRIC ERRORS DETECTION

Polynomial codes are widely used in the construction of automation devices and computing equipment: both in data processing and in the choice of control and diagnostic architecture [23 – 27]. In many applications of these codes,

they use the property of detecting errors in data bits. In [28, 29], the general characteristics of error detection in data vectors by polynomial codes were investigated. Further studies have shown that this class of codes with certain encoders polynomials makes it possible to identify any symmetric and asymmetric errors. Thus, among the polynomial codes multitude, such classes as $SED(m,k)$ and $AED(m,k)$ -codes can be distinguished.

Polynomial codes are constructed as follows:

1. The length of the data vector m is determined.
2. Choose a generating polynomial $G(f)$ with a degree $k = n - m$.
3. Each data vector is written in the form of a polynomial $M(f)$ and multiplied by f^k .
4. The resulting polynomial $f^k M(f)$ is divided into the generating polynomial $G(f)$.
5. The polynomial $R(f)$ corresponding to the remainder of dividing the polynomial $f^k M(f)$ by the encoder polynomial $G(f)$ is represented as a binary number and is written into the check vector.

Most often, the values of residuals are obtained using dividers implemented on shift registers, however, check functions corresponding to the binary form of the values of residuals from division can be obtained directly without using memory circuits. The check functions of polynomial codes are a system of k addition functions modulo two of some part of data bits. The type of polynomial encoder determines the composition of data bits in each check function. The choice of the generative polynomial for a given value of the length of the data vector allows you to build codes with different characteristics of error detection, including obtaining special classes of codes.

In the study of the characteristics of polynomial codes, special types of generating polynomials have been established, which make it possible to obtain $SED(m,k)$ and $AED(m,k)$ -codes. In the course of the analysis, all data vectors were distributed between the check vectors in order to establish the types and error multiplicities of the data vectors for which the check vectors do not change [30]. By such an analysis, special types of polynomials were established, giving codes with certain properties.

Statement 1. *Polynomial codes constructed using generative polynomials of the form:*

$$x^k + x^{k-1} + \dots + x^{k-j} + \dots + x^1 + x^0, \quad (1)$$

where $k=m-1$, $j=\{1;2;\dots\}$, $j < k$.

with an odd number of terms and under the condition $k=m-1$, they detect any symmetric errors.

We give an example of a polynomial code satisfying the condition of statement 1. Take as a generator a polynomial $x^3 + x^2 + x^0$. In Tabl. 1 shows the distribution of all data vectors between the check vectors for a given code.

The analysis of the columns of Tabl. 1 confirms that any symmetric distortions in the data vectors in the code under

consideration are detected since there are no data vectors with the same weight in the same group.

TABLE I. THE POLYNOMIAL CODE WITH THE GENERATOR POLYNOMIAL $x^3 + x^2 + x^0$

Check vectors							
000	001	010	011	100	101	110	111
Data vectors							
0000	0111	0011	0100	0110	0001	0101	0010
1101	1010	1110	1001	1011	1100	1000	1111

Studies also show that for each specific polynomial code, which is the $SED(m,k)$ -code, there are errors in the class of undetectable errors with only one specific multiplicity. For example, for the example in question, these are triple errors, which may be unidirectional or asymmetric.

Statement 2. Polynomial codes constructed using generative polynomials of the form:

$$x^k + x^0, \quad (2)$$

where $k = m-1$,

any asymmetric errors are detected.

Tabl. 2 shows an example of a polynomial code satisfying condition 2.

TABLE II. POLYNOMIAL CODE WITH THE GENERATOR POLYNOMIAL $x^3 + x^0$

Check vectors							
000	001	010	011	100	101	110	111
Data vectors							
0000	0001	0010	0011	0100	0101	0110	0111
1001	1000	1011	1010	1101	1100	1111	1110

It should be noted that polynomial $AED(m,k)$ -codes have the property of detecting any errors in data vectors, except for twofold ones. This property is traced for all polynomial codes with generating polynomials of the form (2).

In [28, 29] it is shown that polynomial codes, which form polynomials that do not contain a free member, do not detect a large number of errors, some of them are not noise-resistant codes (one-time distortions are not detected), some check vectors are not used, etc. Therefore, only polynomial codes, the polynomials of which have a free member, are promising for solving problems in the organization of reliable automation systems and computing equipment. Tabl. three lists all the polynomial codes that can be applied in the problems under consideration, also those codes for which the conditions of statements 1 and 2 are fulfilled.

VI. EMPIRICAL DATA

Experiments were conducted to detect errors at the outputs of benchmarks from the LGSynth'89 set to confirm the properties of polynomial codes [31]. Circuits in this set are defined in various ways, including in the form of sheets describing topology (net-lists). This allows you to analyze their work when making various kinds of faults in the internal structure. Faults manifest themselves as logical error signals and propagate along paths leading to the outputs of benchmarks, distorting the values on them. In an experiment, the

model of constant faults at the exits of logic gates of an inner pattern of the combinational scheme was selected. As test circuits schemes with a small number of outputs ($m = 3, 4, 5$) were selected and simulated all single constant faults in their structures are made. Then, the total number of detected and undetectable errors in the testing of benchmarks by a certain polynomial code is recorded.

TABLE III. GENERATOR POLYNOMIALS THAT FORM SOME "SPECIAL" POLYNOMIAL CODES

Polynomial	The number of check bits	Special code class
$x^2 + x^0$	2	$AED(3,2)$ -code
$x^2 + x^1 + x^0$	2	$SED(3,2)$ -code
$x^3 + x^0$	3	$AED(4,3)$ -code
$x^3 + x^1 + x^0$	3	$SED(4,3)$ -code
$x^3 + x^2 + x^0$	3	$SED(4,3)$ -code
$x^3 + x^2 + x^1 + x^0$	3	–
$x^4 + x^0$	4	$AED(5,4)$ -code
$x^4 + x^1 + x^0$	4	$SED(5,4)$ -code
$x^4 + x^2 + x^0$	4	$SED(5,4)$ -code
$x^4 + x^2 + x^1 + x^0$	4	–
$x^4 + x^3 + x^0$	4	$SED(5,4)$ -code
$x^4 + x^3 + x^1 + x^0$	4	–
$x^4 + x^3 + x^2 + x^0$	4	–
$x^4 + x^3 + x^2 + x^1 + x^0$	4	$SED(5,4)$ -code
$x^5 + x^0$	5	$AED(6,5)$ -code
$x^5 + x^1 + x^0$	5	$SED(6,5)$ -code
$x^5 + x^2 + x^0$	5	$SED(6,5)$ -code
$x^5 + x^2 + x^1 + x^0$	5	–
$x^5 + x^3 + x^0$	5	$SED(6,5)$ -code
$x^5 + x^3 + x^1 + x^0$	5	–
$x^5 + x^3 + x^2 + x^0$	5	–
$x^5 + x^3 + x^2 + x^1 + x^0$	5	$SED(6,5)$ -code
$x^5 + x^4 + x^0$	5	$SED(6,5)$ -code
$x^5 + x^4 + x^1 + x^0$	5	–
$x^5 + x^4 + x^2 + x^0$	5	–
$x^5 + x^4 + x^2 + x^1 + x^0$	5	$SED(6,5)$ -code
$x^5 + x^4 + x^3 + x^0$	5	–
$x^5 + x^4 + x^3 + x^1 + x^0$	5	$SED(6,5)$ -code
$x^5 + x^4 + x^3 + x^2 + x^0$	5	$SED(6,5)$ -code
$x^5 + x^4 + x^3 + x^2 + x^1 + x^0$	5	–

Tabl. 4 shows the results of experiments with selected benchmarks.

The experimental results confirm the correctness of the properties of polynomial codes established in the paper. In this case, it is possible to turn on several important features of the application of polynomial codes at the testing of combinational benchmarks. First, despite the fact that the polynomial $SED(m,k)$ -codes do not belong to the class and $AED(m,k)$ -codes, they detect a significant proportion of asymmetric errors at the outputs of benchmarks, and for

some circuits – all asymmetrical errors. A similar pattern is inherent for polynomial $AED(m,k)$ -codes with respect to symmetric errors. Secondly, even though polynomial codes in the undetectable class have unidirectional errors, they detect a significant proportion of unidirectional errors at the outputs of benchmarks, and for some variants all unidirectional errors. Thirdly, for several benchmarks, the use of polynomial codes turns out to be justified when any types of errors are detected, and the results are comparable to the use of the duplication method in monitoring. However, this result is explained by the high redundancy of polynomial codes and the presence of $k=m-1$ check bit in them. According to this indicator, polynomial codes can be compared with codes weighted by degrees of the number of two transitions between bits that occupy adjacent positions in the data vectors used in organizing concurrent checking systems for combinational circuits [32].

V. CONCLUSION

Among a variety of polynomial codes, the codes formed by special polynomials, allowing to find in data vectors of codes separately any symmetrical and any asymmetric errors can be selected. Such codes form the classes $SED(m,k)$ and $AED(m,k)$ -codes which can be applied at the automation systems design with fault detection. Failure monitoring of a device is carried out based on whether the belonging of errors to the classes of symmetrical or asymmetric errors.

In that case, subjects to diagnosing or should have the structures allowing only symmetric (or only asymmetric) errors, or the corresponding groups of outputs should be highlighted on the set of their outputs. The method of organization of check systems in this case is similar to the method of searching for unidirectionally independent groups using $UED(m,k)$ -codes.

REFERENCES

- [1] V. Kharchenko, Yu. Kondratenko, and J. Kacprzyk "Green IT Engineering: Concepts, Models, Complex Systems Architectures", Springer Book ser. "Studies in Systems, Decision and Control", Vol. 74, 2017, 305 p.
- [2] V. Hahanov "Cyber-Physical Computing for IoT-driven Services", N.Y., Springer Int. Pub. AG, 2018, 279 p.
- [3] V. Kuntsevich, V. Gubarev, Yu. Kondratenko, D. Lebedev, and V. Lysenko "Control Systems: Theory and Applications", River Pub. Ser. in Automation, Control and Robotics, 2018.
- [4] R. Ubar, J. Raik, and H.-T. Vierhaus "Design and Test Technology for Dependable Systems-on-Chip (Premier Reference Source)", Inf. Science Reference, Hershey – N.Y., IGI Global, 2011, 578 p.
- [5] Z. Navabi "Digital System Test and Testable Design: Using HDL Models and Architectures", Springer Science+Business Media, LLC 2011, 435 p.
- [6] E. Fujiwara "Code Design for Dependable Systems: Theory and Practical Applications", John Wiley & Sons, 2006, 720 p.
- [7] P.K. Lala "Principles of Modern Digital Design", N.-J.: John Wiley & Sons, 2007, 436 p.

TABLE IV. EMPIRICAL DATA

№	Bench. name	Number inputs / outputs	Total number errors by type			Polynomial	Total number undetectable errors by type		
			Unidirectional	Symmetrical	Asymmetrical		Unidirectional	Symmetrical	Asymmetrical
1	cm82a	5 / 3	0	68	4	$x^2 + x^0$	0	0	0
						$x^2 + x^1 + x^0$	0	0	4
2	cm85a	11 / 3	0	176	0	$x^2 + x^0$	0	48	0
						$x^2 + x^1 + x^0$	0	0	0
3	b1	3 / 4	0	2	0	$x^3 + x^0$	0	0	0
						$x^3 + x^1 + x^0$	0	0	0
						$x^3 + x^2 + x^0$	0	0	0
4	cmb	16 / 4	39456	6	0	$x^3 + x^0$	0	0	0
						$x^3 + x^1 + x^0$	0	0	0
						$x^3 + x^2 + x^0$	0	0	0
5	z4ml	7 / 4	0	128	32	$x^3 + x^0$	0	0	0
						$x^3 + x^1 + x^0$	0	0	0
						$x^3 + x^2 + x^0$	0	0	0
6	cm162a	14 / 5	314067	1920	1344	$x^4 + x^0$	224	0	0
						$x^4 + x^1 + x^0$	224	0	0
						$x^4 + x^2 + x^0$	224	0	0
						$x^4 + x^3 + x^0$	224	0	0
						$x^4 + x^3 + x^2 + x^0$	224	0	0
7	cm163a	16 / 5	1203648	10368	7296	$x^4 + x^0$	0	256	0
						$x^4 + x^1 + x^0$	0	0	0
						$x^4 + x^2 + x^0$	0	0	0
						$x^4 + x^3 + x^0$	0	0	128
						$x^4 + x^3 + x^2 + x^0$	32	0	32

- [8] W.E. Ryan, and S. Lin "Channel Codes: Classical and Modern", Cambridge University Press, 2009, 708 p.
- [9] F.F. Sellers, M.-Y. Hsiao, and L.W. Bearnson "Error Detecting Logic for Digital Computers", N.Y.: McGraw-Hill, 1968, XXI + 295 p.
- [10] M. Nicolaidis, and Y. Zorian "On-Line Testing for VLSI – A Compendium of Approaches", JETTA, 1998, Vol. 12, Iss. 1-2, pp. 7-20, doi: 10.1023/A:1008244815697.
- [11] S. Mitra, and E.J. McCluskey "Which Concurrent Error Detection Scheme to Choose?", Proc. of Int. Test Conference, 2000, USA, Atlantic City, NJ, 03-05 October 2000, pp. 985-994, doi: 10.1109/TEST.2000.894311.
- [12] M. Göessel, V. Ocheretny, E. Sogomonyan, and D. Marienfeld "New Methods of Concurrent Checking: Edition 1", Dordrecht: Springer Science+Business Media B.V., 2008, 184 p.
- [13] V.V. Sapozhnikov, VI.V. Sapozhnikov, and D.V. Efanov "Classification of errors in the information vectors of systematic codes" (in Russ.), Izvestiya vuzov. Instr. making, 2015, Vol. 58, Iss. 5, pp. 333-343, doi: 10.17586/0021-3454-2015-58-5-333-343.
- [14] V.V. Sapozhnikov, VI.V. Sapozhnikov, and D.V. Efanov "Codes with the summation, detecting any symmetric errors" (in Russ.), Electronic Modelling, 2017, Vol. 39, Iss. 3, pp. 47-60.
- [15] E.S. Sogomonyan, and M. Gössel "Design of Self-Testing and On-Line Fault Detection Combinational Circuits with Weakly Independent Outputs", JETTA, 1993, Vol. 4, Iss. 4, pp. 267-281.
- [16] F.Y. Busaba, and P.K. Lala "Self-Checking Combinational Circuit Design for Single and Unidirectional Multibit Errors", JETTA, 1994, Vol. 5, Iss. 1, pp. 19-28.
- [17] V.V. Saposhnikov, A. Morosov, VI.V. Saposhnikov, and M. Göessel "A New Design Method for Self-Checking Unidirectional Combinational Circuits", JETTA, 1998, Vol. 12, Iss. 1-2, pp. 41-53, doi: 10.1023/A:1008257118423.
- [18] A. Morosow, V.V. Sapozhnikov, VI.V. Sapozhnikov, and M. Goessel "Self-Checking Combinational Circuits with Unidirectionally Independent Outputs", VLSI Design, 1998, Vol. 5, Iss. 4, pp. 333-345, doi: 10.1155/1998/20389.
- [19] D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "Conditions for Detecting a Logical Element Fault in a Combination Device under Concurrent Checking Based on Berger's Code", Autom. Remote Control, 2017, Vol. 78, Iss. 5, pp. 891-901, doi: 10.1134/S0005117917040113.
- [20] V. Sapozhnikov, VI. Sapozhnikov, and D. Efanov "Search Algorithm for Fully Tested Elements in Combinational Circuits, Controlled on the Basis of Berger Codes", Proc. of 15th EWDTS, Novi Sad, Serbia, September 29 – October 2, 2017, pp. 99-108, doi: 10.1109/EWDTS.2017.8110085.
- [21] D. Das, and N.A. Touba "Synthesis of Circuits with Low-Cost Concurrent Error Detection Based on Bose-Lin Codes", JETTA, 1999, Vol. 15, Iss. 1-2, pp. 145-155, doi: 10.1023/A:1008344603814.
- [22] A.A. Blyudov, D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "On Codes with Summation of Unit Bits in Concurrent Error Detection Systems", Autom. Remote Control, 2014, Vol. 75, Iss. 8, pp. 1460-1470, doi: 10.1134/S0005117914080098.
- [23] S. Bayat-Sarmadi, and M.A. Hasan "Concurrent Error Detection of Polynomial Basis Multiplication Over Extension Fields Using a Multiple-Bit Parity Scheme", 20th IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems (DFT 2005), 3–5 October 2005.
- [24] Bayat-Sarmadi S., Hasan M.A. "On Concurrent Detection of Errors in Polynomial Basis Multiplication", IEEE Transactions on VLSI Systems, 2007, vol. 15, pp. 413-426, doi: 10.1109/TVLSI.2007.893659.
- [25] Qiu W., Zhang X., Li H., Wang Z., Zhang Y., Zheng Z. "Concurrent All-Cell Error Detection in Semi-Systolic Multiplier Using Linear Codes", Appl. Math. & Inform. Sciences, 2013, Vol. 7, No 3, pp. 947–954.
- [26] El-Khamy M., Lee J., Kang I. "Detection Analysis of CRC-Assisted Decoding", IEEE Comm. Letters, 2015, Vol. 19, Issue 3, pp. 483-486.
- [27] D. Gangopadhyay, and A. Reyhani-Masoleh "Multiple-Bit Parity-Based Concurrent Fault Detection Architecture for Parallel CRC Computation", IEEE Trans. on Computers, 2016, Vol. 65, Issue 7, pp. 2143-2157.
- [28] V.V. Sapozhnikov, VI.V. Sapozhnikov, D.V. Efanov, and R.B. Abdullaev "On the properties of Polynomial Codes in CED Systems" (in Russ), Informatics and control systems, 2018, Iss. 2, pp. 50-61, doi: 10.22250 / isu.2018.56.50-61.
- [29] D. Efanov, D. Plotnikov, V. Sapozhnikov, VI. Sapozhnikov, and R. Abdullaev "Experimental Studies of Polynomial Codes in Concurrent Error Detection Systems of Combinational Logical Circuits", Proc. of 16th EWDTS, Kazan, Russia, September 14-17, 2018, pp. 184-190, doi: 10.1109/EWDTS.2018.8524684.
- [30] D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "On Summation Code Properties in Functional Control Circuits", Autom. and Remote Control, 2010, Vol. 71, Iss. 6, pp. 1117-1123, doi: 10.1134/S0005117910060123.
- [31] "Collection of Digital Design Benchmarks", [http://ddd.fit.cvut.cz/prj/Benchmarks/].
- [32] V.V. Sapozhnikov, VI.V. Sapozhnikov, D.V. Efanov, and V.V. Dmitriev "New Structures of the Concurrent Error Detection Systems for Logic Circuits", Autom. Remote Control, 2017, Vol. 78, Iss. 2, pp. 300-312, doi: 10.1134/S0005117917020096.

Diagnostics of Audio-Frequency Track Circuits in Continuous Monitoring Systems for Remote Control Devices: Some Aspects

Dmitrii V. Efanov,
DSc, Professor at “Automation,
Remote Control and
Communication
on Railway Transport”,
Russian University of Transport
(MIIT),
Moscow, Russia
TrES-4b@yandex.ru

German V. Osadchy,
Technical Director of Scientific
and Technical Center
“Integrated Monitoring
Systems” LLC,
St. Petersburg, Russia
osgerman@mail.ru

Valerii V. Khóroshev,
PhD Student, Department
of “Automation, Remote
Control and Communication on
Railway Transport”, Russian
University of Transport (MIIT),
Moscow, Russia
Hvv91@icloud.com

Dmitrii A. Shestovitskiy
PhD, Associate Professor,
Department of “Bridges”,
Emperor Alexander I
St. Petersburg State
Transport University,
St. Petersburg, Russia
diamond0110@mail.ru

Abstract—The audio-frequency track circuits are the main sensor for monitoring the position of railway transport on the Russian railways. Failures of railway track circuits constitute about 30% of the failures of all railway automation and remote-control equipment. The paper analyzes the features of the technical condition continuous monitoring for the audio-frequency track circuits by means of the Hardware-software complex of dispatch control. An algorithm for processing diagnostic information is proposed for organizing of track circuits monitoring in an automatic mode considering the states of the main track circuit units, such as a frequency generator, track receiver and track relay. As well as measured analogue values at test points. In this paper, the main features of various diagnostic situations were established, including the failure and pre-failure condition based on the expert response. This allowed the creation of an intelligent decision support subsystem for the situational centres technical staff. We used the methods of the theory of discrete devices, information and coding theory, as well as technical diagnostics.

Keywords—*railway automation and remote control, health monitoring, audio-frequency track circuit, pre-failure condition, analysis automation*

I. INTRODUCTION

The train traffic control organization on railways use a complex of automation and remote-control devices [1–4]. Automation devices perform critical technological operations to train control algorithms implementation. They are built in compliance with all safety requirements [5]. However, failures are not excluded during the operation of automation devices. The probability of failure is higher for devices located near the railroad tracks. The unreliability of railway automation device trackside objects is caused by their location and interaction with rolling stock [6–8]. As a result, over 80% of the railway automation device failures account for trackside equipment.

To increase the reliability and ensure the fault tolerance of trackside automation devices, which include switch points, railway track circuit equipment, automatic level crossing equipment, etc., they carry out their periodic maintenance. Modern means of automatic diagnosis and monitoring of the technical condition are used [9]. Such systems have been actively developed since the end of the 20th century and are integrated into train control systems. These

systems are mainly present measuring controller and measuring sensors [10–17]. According to the measuring equipment, it is possible to estimate the state of the trackside equipment.

On the ex-USSR countries territory, the continuous monitoring systems of automation equipment are widespread. This is a separate class of devices [9]. They represent the means of external technical diagnostics. Their measuring controllers are connected to the inspection point of railway automation circuits. They collect diagnostic data in real time with a predetermined period and send it to information hubs, where it is processed and issued to technical personnel of situational monitoring centres. The use of continuous monitoring systems allows technicians to quickly respond to the occurrence of abnormal situations in the event of a pre-failure condition, which helps to prevent failures. Thus, such important property of control systems as fault tolerance is supported.

The authors of the paper pay attention to the problem and features of monitoring of the tone frequency track circuits.

II. OBTAINING DATA ON THE STATE OF RAIL CIRCUITS

The audio-frequency track circuits are widespread on the railways of the Russian Federation and are used on various lines, including speed and high-speed traffic. These devices operate in conditions of external destructive factors in the open air and are also subject to electromagnetic effects and interference from rolling stock [6, 9].

The audio-frequency track circuits operate at carrier frequencies of 420, 480, 580, 720, 780 Hz modulated by frequencies of 8 or 12 Hz. Almost all the equipment is located at the signalling box or in a transportable module. The exceptions are the track transformers, balancer and resistors, through which the signal box equipment is connected to the rails. They are in travel boxes. The signal box equipment includes track generators (TG), track filters (TF), track receivers (Trec) and track relays (TR). The track generator connects to the feed-end through the track filter. The generator is powered by alternating current with a frequency of 50 Hz and a voltage of 35 V. It produces the required frequency in the range of 420 – 780 Hz, which is required for operation. On the receiver (relay) ends are track receiver and track relay. In the absence of rolling stock, the track receiver

tuned to the appropriate frequency is in the state. The track relay is energized and transmits information about the vacancy of the monitored track section. When the track circuits are shunting by a wheel pair, damage to the rail line, the track receiver and the track relay are turned off. The condition of the track circuit in this case in the train control system is interpreted as occupied. The track receiver is powered by a current of 50 Hz and a voltage of 17.5 V.

In order to organize the monitoring of the track circuit technical condition, the main operating parameters of the equipment are monitored using specialized diagnostic controllers. These are measuring controllers of the audio-frequency track circuits of various modifications [9]. In Fig. 1 shows an example of a rail track controller. In Fig. 2 is a diagram of connecting the controller to the monitoring object.

The measurement controller is connected in parallel to the track relay, the push receiver, the track generator and the track filter.



Fig. 1. Measuring controller for the audio-frequency track circuits.

In order to prevent the dangerous influence of the measuring controller on the circuit of the audio-frequency track circuit, the connection is made through the block of protective resistors R1-R2, R3-R4, ..., R9-R10. Each resistor has a resistance of 6.81 kΩ and makes it possible to eliminate any

interfering influence of the diagnostic equipment on the rail circuit itself.

The controller has eight channels of obtaining analogue information and allows you to measure voltage with a polling period of 8–12 second. If the controller MK-8 is used, then its modification is used to connect to different points of the track circuit. The MK-8 controller is connected in parallel to the track receiver and measures the AC voltage in the range of 0–2 V. MK-8-01 connects to the track generator and measures the voltage in the range of 0–12 volts. MK-8-02 measures the data from the track relay.

The controllers are universal and can be connected to any of the nodes indicated above. Among the advantages are an increased measurement speed (2–5 second), high-noise immunity and advanced self-diagnostics functions of the devices.

III. DIAGNOSTIC INFORMATION INTRODUCTION

Diagnostic devices measure the voltage at the control points of the audio-frequency track circuits. Conduct primary data processing and transmit them to the diagnostic information concentrator located on the signalling box. From the hub, the data is sent to the service personnel automated workplace, via the data transfer channel to the central hub, to the situational centres and to the management of JSC «Russian Railways».

In the continuous monitoring system "Hardware-software complex dispatch control" diagnostic information is presented in graphical form. For each track circuit in the automated workplace operator interface, graphs of voltage changes at the control points of the track circuit are derived. Diagram can be displayed both in a separate form for each monitoring point (Fig. 3) and in a combined one (Fig. 4).

Carrying out the analysis of graphic data, the monitoring technologist makes a conclusion about the technical condition of the track circuit and issues informational messages to the service personnel.

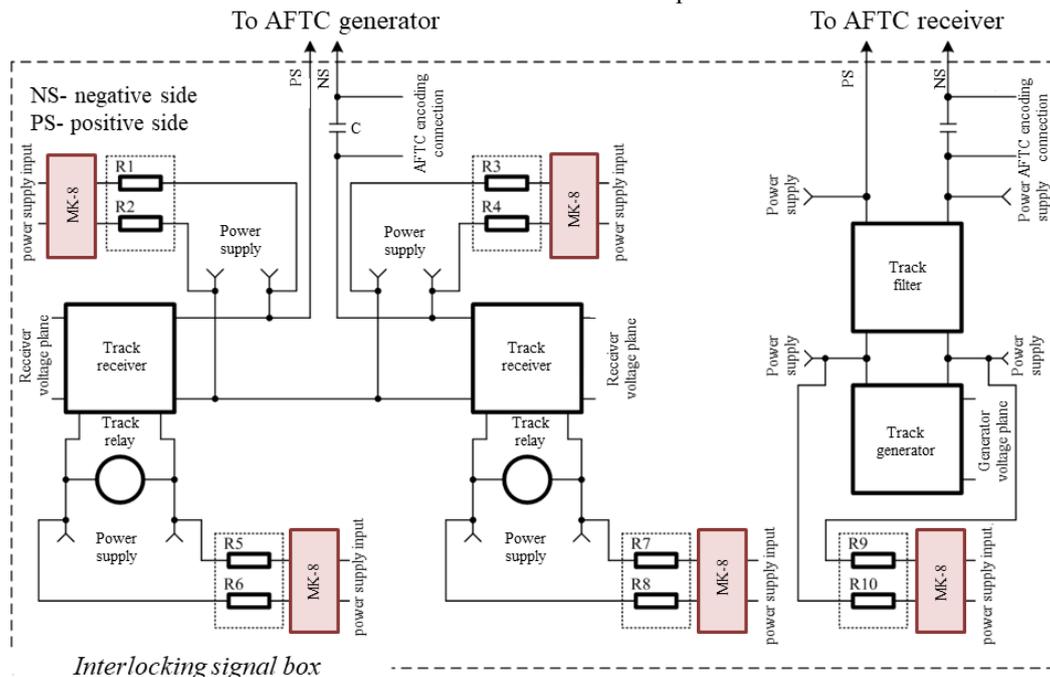


Fig. 2. Scheme of connecting the measuring controller.

For example, data analysis in Fig. 3 and Fig. 4 indicates that during normal operation of the generator (the upper graph of Fig. 3), voltage tracing is observed at the track receivers of the ends A and B of the track circuit (the second and third graphs of Fig. 3 above). This leads to the occasional disengagement the track relays of the two ends (the two lower graphs of Fig. 3). This diagnostic situation corresponds to the pre-failure state of the track receiver.

A similar analysis is carried out manually, and the human factor influences the monitoring process itself. Technologist or maintenance personnel is able to skip the development of a malfunction, which, ultimately, will lead to the failure of automation equipment.

In the course of research of many graphs of technological situations related to the operation of audio-frequency

track circuits, features were found inherent in various types of failure or pre-failure situations. Below gives some examples of the most common occurrences of audio-frequency track circuits failures and pre-failures.

Figures 6-20 show voltage diagrams at measuring points for a typical technological situation (failure or pre-failure). For these figures, a description and a brief analysis is provided, which will help monitoring staff to more easily analyze changes in the values obtained from measuring controllers of audio-frequency track circuits. With further improvement of the monitoring technology, the logic of a technologist work for analyzing diagnostic information should form the basis of data processing by software methods.

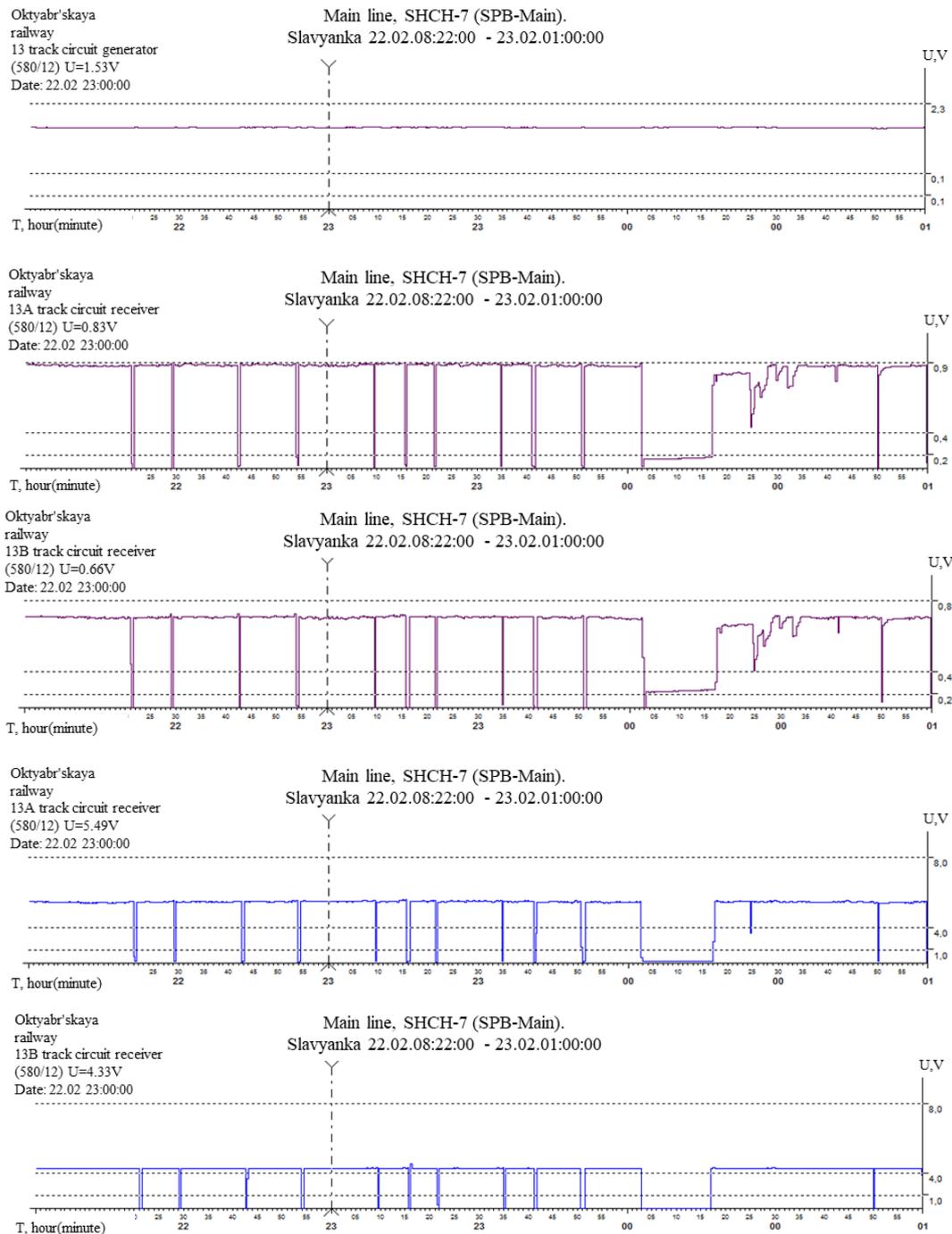


Fig. 3. The measured voltage values in a separate form.

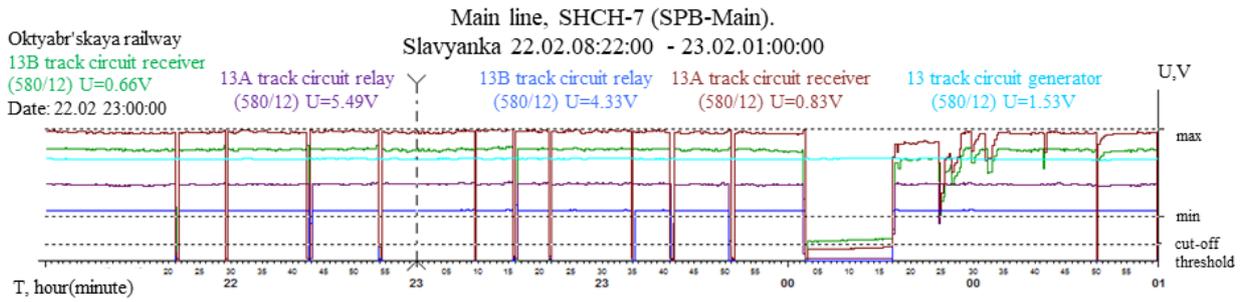


Fig. 4. The measured voltage values in a combined form.

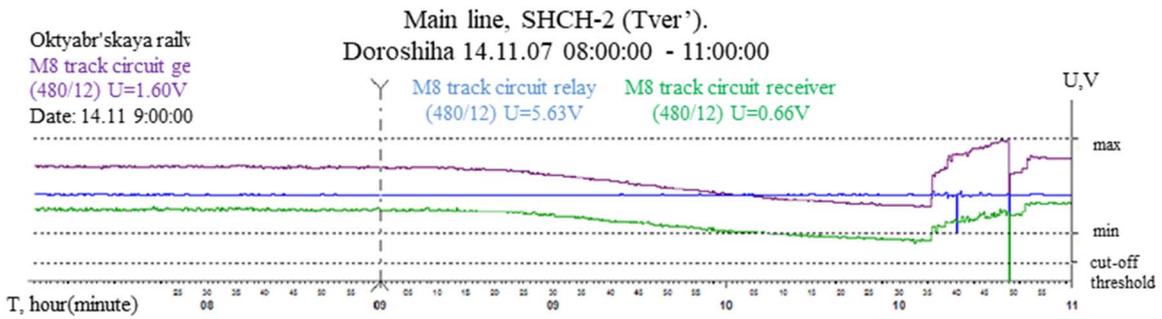


Fig. 5. Track generator pre-failure.

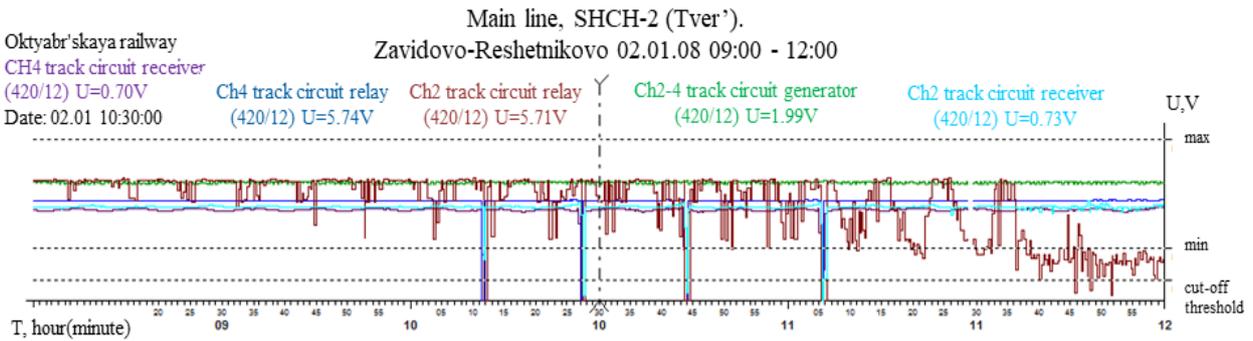


Fig. 6. Track receiver pre-failure.

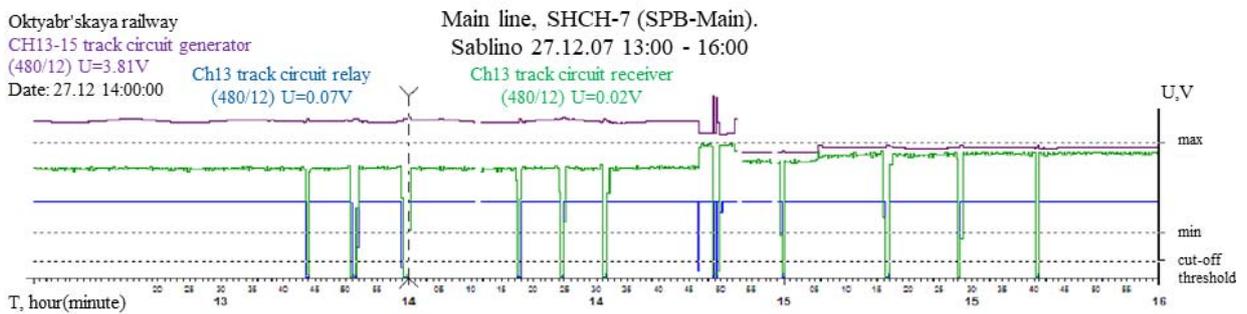


Fig. 7. Track filter pre-failure.

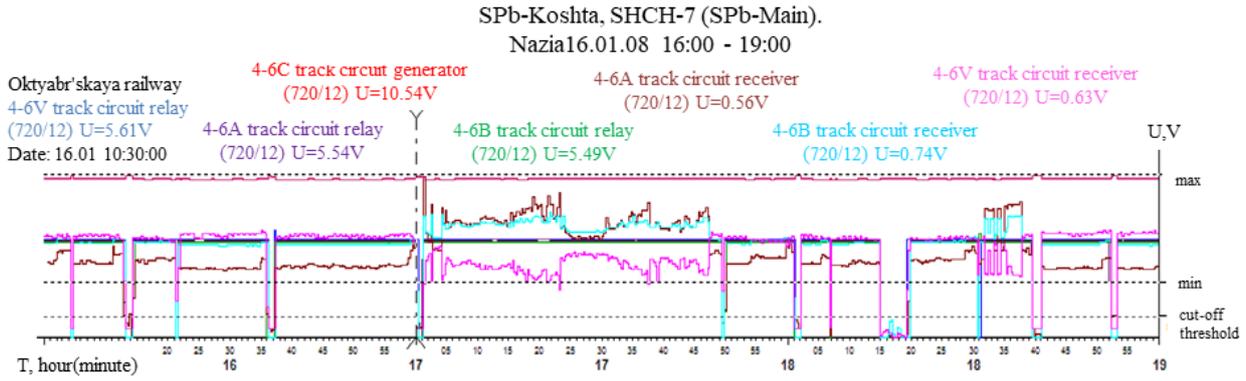


Fig. 8. Probable pre-failure of audio-frequency track circuits trackside equipment.

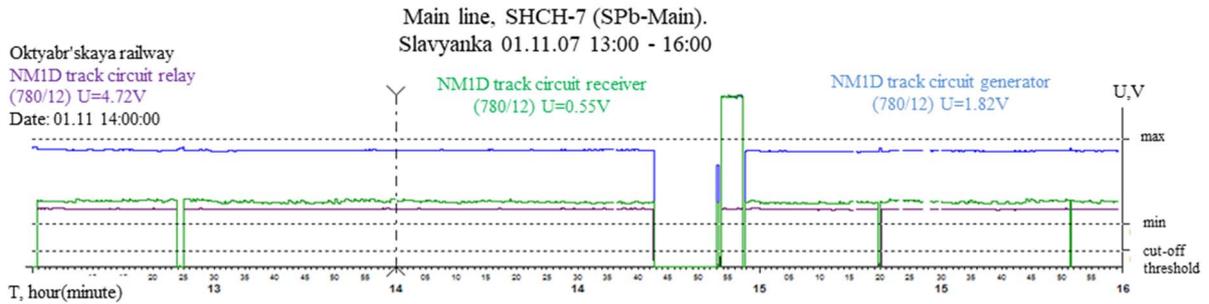


Fig. 9. Track generator failure.

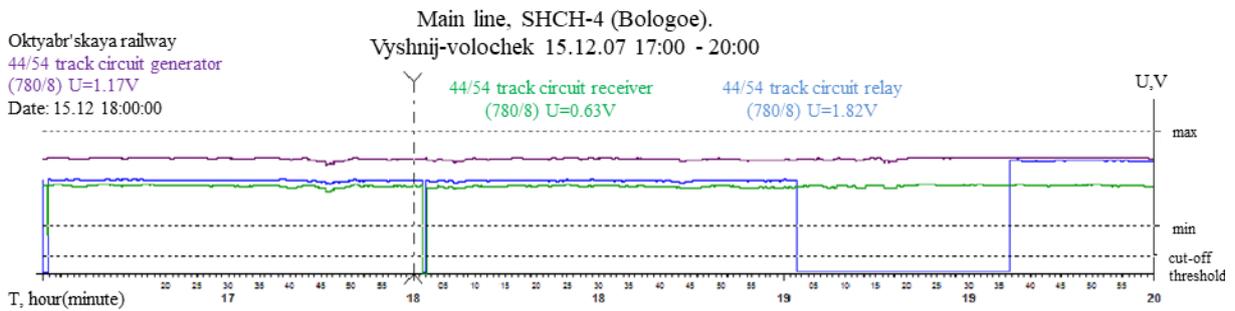


Fig. 10. Track receiver failure.

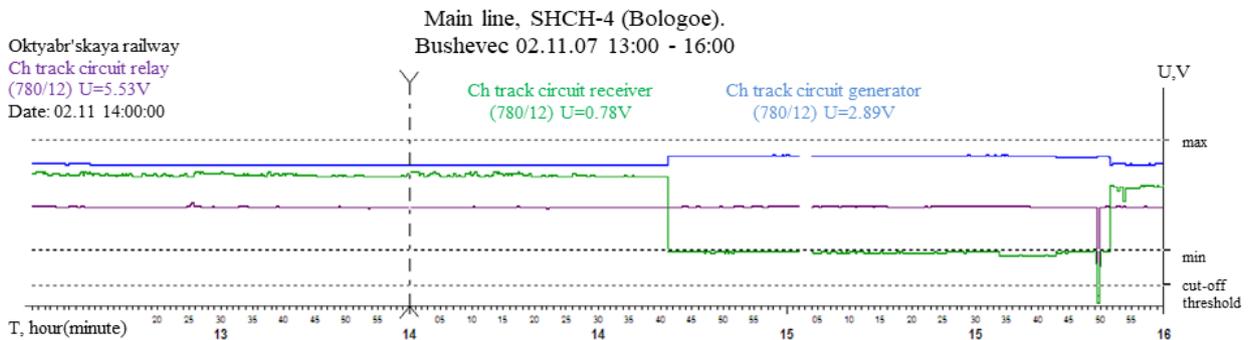


Fig. 11. Track filter failure.

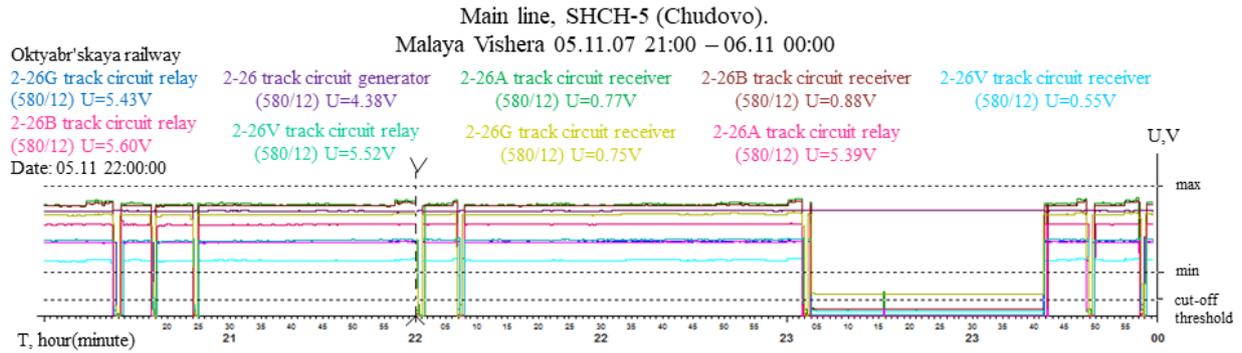


Fig. 12. Failure of audio-frequency track circuits trackside equipment.

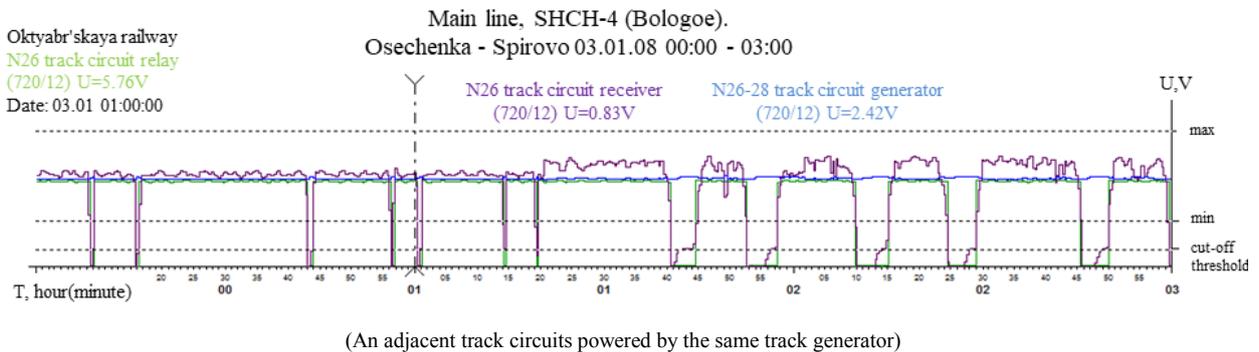
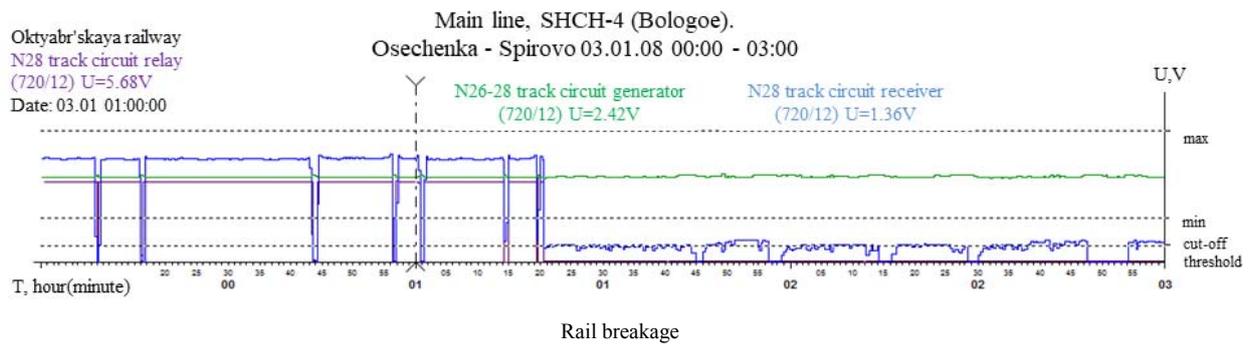


Fig. 13. Rail breakage.

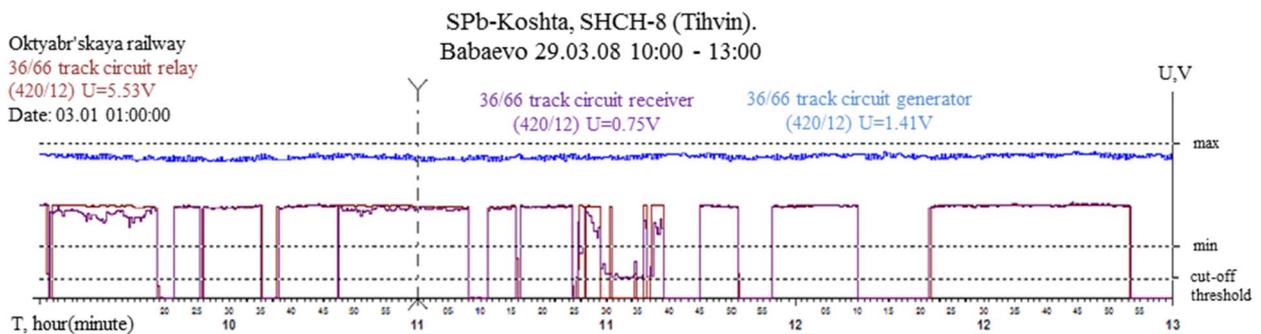


Fig. 14. The intermittent fault of audio-frequency track circuits trackside equipment.

In the pre-failure state of the track generator on Fig. 5, a slow voltage drop will be observed at its output. It will also observe a voltage drop at the input of the track receiver. It is also possible to pulsate voltage parameters between the upper and lower limits.

When analyzing disturbances in the track receiver (Fig. 6), the following distinctive features are observed in the voltage diagram:

1. The voltage at the output of the track generator remains unchanged.
2. The voltage at the input of the track receiver ripple within the lower and upper limits.
3. The voltage at the output of the track generator is almost unchanged.

If a branched track circuit is being analyzed, then if one of the branch receivers fails, other receivers may experience a slight increase in voltage. be observed on other receivers.

In the pre-failure state of the track filter on Fig. 7, the voltage at the output of the track generator increases by 5 – 10%. In this case, the voltage at the receiver input may decrease to the minimum value.

Possible pre-failure conditions in trackside equipment (Fig.8):

- increase in transient resistance in cable and butt connectors;
- the growth of transient resistance in the track boxes and cable terminal coupling;
- short circuit in track circuits.

With these pre-requisitions, the voltage at the input of the track receiver decreases with a ripple within the normal range.

To localize the fault, you can use the measured values of the voltages on the track receiver of all the branches of the track circuit.

When a track generator fails (Fig. 9), a sharp voltage drop occurs at its output. At the entrance and exit of the track receiver, the voltage drops below normal.

When a track receiver fails, as on Fig. 10. The sharp decrease in voltage at its output is observed, while the voltage diagram at the input of the track receiver and the output of the track generator remain almost unchanged. For these features, it can be concluded that the voltage from the output of the track receiver does not flow to the track relay, and the voltage is present at the input of the track receiver and is normal.

If the track filter fails (Fig. 11), a slight increase in voltage will be observed at the output of the track generator, and the voltage at the input and output of the track receiver will drop below the normal.

Possible state failures in trackside equipment (Fig. 12):

- increase in transient resistance and breaks in cable and butt connectors;
- the growth of transient resistance and breaks in the track boxes and cable terminal;
- short circuit in rail circuits.

Failure may cause a voltage drop at the input and output of the receiver to a level below the norm. The voltage on the track generator is almost unchanged.

To localize the fault, you can use the measured values of the voltages on the track receiver of all the branches of the track circuit.

After analyzing the voltage diagram of the track circuit on Fig. 13, in which a rail breakage was detected, the following distinguishing features were identified:

1. At the output of the track generator at which the fracture occurred, the voltage does not change.
2. At the entrance of the track receiver, where the fracture occurred, the voltage drops below the minimum value and a weak ripple is observed (the closer the break to the feed end of the track circuit, the higher the amplitude and frequency of ripple).
3. At the output of the track receiver, the voltage does not drop to zero.

If the track circuit has an adjacent one that is powered by the same generator, then the following manifestations are observed on it:

1. At the output of the track generator, a small voltage increases of 5 – 10%.
2. At the input of the track receiver, a voltage rises of 5 – 10% occurs.

Possible intermittent failures in trackside equipment (Fig. 14):

- short circuit of the insulating joint;
- transient resistances and breaks in cable and butt connectors;
- transient resistances and breaks in track boxes;
- short circuit in the track circuit.

With an intermittent failure in the possible voltage drop at the receiver input to a level below the norm and its ripple from the maximum value and below the minimum. In this case, the voltage at the receiver output tends to zero.

To localize the fault, you can use the data:

1. Voltage track receiver on all branches.
2. In the event of a short circuit of an insulating joint, parameters of adjacent track circuits can be used.

The main condition for the operation of the algorithm is almost constant voltage value at the output of the track generator $U_{tg} \approx \text{const}$ since the voltage at the track receiver depends on the voltage change on the track generator. Suppose that this condition is fulfilled, and the voltage of the track generator corresponds to the norm set for this track circuit (logical operator $\langle 1 \rangle$). At the same time, it is also necessary to control the vacancy of adjacent track circuits for the lack of occupancy by their mobile units.

The algorithm must operate continuously in time, which means that feedback is required to create a data processing cycle.

The choice of the U_{max} will depend on the type of curve of the schedule, the distance of the device from the location

of the staff, the climate zone and the work schedule of the staff.

IV. AUTOMATING DIAGNOSTIC ANALYSIS ALGORITHM

Many diagnostic situations for each component of the track circuit were described by the authors in the diagnostic model and adapted to the software of "Hardware-software complex dispatch control". In Fig. 15 shows an algorithm for analyzing the pre-failure states of a track receiver, compiled using an appropriate diagnostic model.

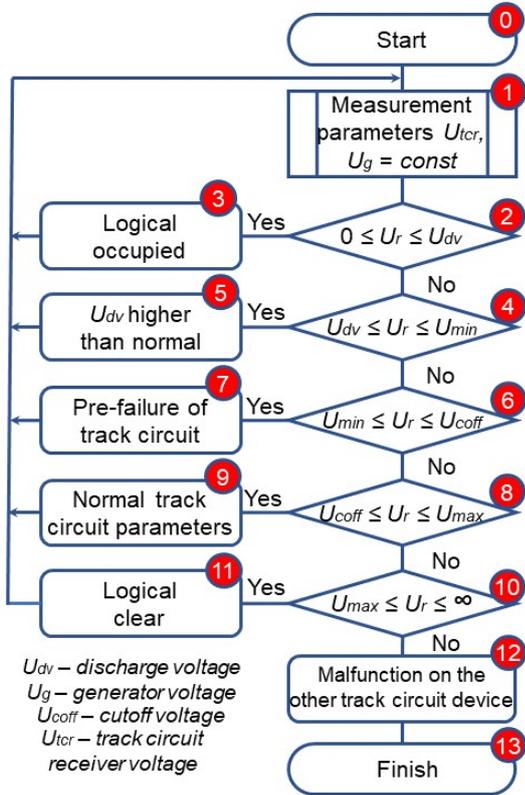


Fig. 15. Algorithm for fixing the pre-failure state.

V. CONCLUSION

Continuous monitoring of automation device parameters is extremely important. Monitoring systems for automation equipment complement the known means of maintaining the resiliency of the entire railway infrastructure and rolling stock [18–22] and allow for the safe and non-stop passage of trains along railways. The organization of continuous monitoring of the audio-frequency track circuits and the automation of diagnostic information processing is an important step towards improving the reliability of operation and the possibility of preventing malfunctions at the stage of development of their pre-failure conditions. Conducted research and developed catalogues of pre-failure and abandoned states allow in practice to simplify the analysis of technological situations arising in the operation of equipment. In addition, based on the establishment of typical forms of diagrams with subsequent machine analysis, it is possible to automate the processes of identifying failures and pre-failures.

Using the presented approach in the monitoring systems, it is possible to achieve a high level of response to the development of faults in the existing train control systems.

REFERENCE

- [1] M.J. Morley "Safety-Level Communication in Railway Interlockings", Science of Computer Programming, 1997, vol. 29, pp. 147-170.
- [2] T. Takashige "Signalling Systems for Safe Railway Transport", Japan Railway & Transport Review 21, 1999, pp. 44-50.
- [3] G. Theeg, and S. Vlasenko "Railway Signalling & Interlocking: 2nd Edition", Germany, Hamburg: PMC Media House GmbH, 2018, 458 p.
- [4] U. Yildirim, M.S. Durmuş, and M.T. Söylemez "Fail-Safe Signalization and Interlocking Design for a Railway Yard: An Automation Petri Net Approach", Proceedings of 7th International Symposium on Intelligent and Manufacturing Systems (IMS 2010), Sarajevo, Bosnia Herzegovina, September 15-17, 2010, pp. 461-470.
- [5] D.V. Gavzov, V.V. Sapozhnikov, and V.I. Sapozhnikov "Methods for Providing Safety in Discrete Systems", Automation and Remote Control, 1994, vol. 55, issue 8, pp. 1085-1122.
- [6] V.I. Shamanov "The Magnetic Properties of Rail Lines and Levels of Interferences for the Apparatus of Automatic Control and Remote-control", Russian Electrical Engineering, 2015, vol. 86, issue 9, pp. 548-552, doi:10.3103/S1068371215090102.
- [7] V.I. Shamanov "Alternating Traction Current Dynamics in Track Lines of Double-Track Hauls", Russian Electrical Engineering, 2016, vol. 87, issue 10, pp. 566-571, doi: 10.3103/S1068371216100060.
- [8] P.F. Bestem'yanov "Energy-Efficient Algorithms for Assessment of the Rail-Circuit Operation", Russian Electrical Engineering, 2017, vol. 88, issue 9, pp. 557-562, doi: 10.3103/S106837121709005X.
- [9] D.V. Efanov "Concurrent Checking and Monitoring of Railway Automation and Remote Control Devices" (in Russ.), St. Petersburg, Emperor Alexander I St. Petersburg state transport university, 2016, 171 p.
- [10] F.B. Zhou, M.D. Duta, M.P. Henry, S. Baker, and C. Burton "Remote Condition Monitoring for Railway Point Machine", 2002 ASME/IEEE Joint Railroad Conference, 23-25 April 2002, Washington, DC, USA, doi: 10.1109/RRCON.2002.1000101.
- [11] J.L.M. Domingues "Diagnostic Levels in Railway Applications", Signal + Draht, 2004, Issue 1/2, pp. 31-34.
- [12] O. F. Eker, F. Camci, A. Guclu, H. Yilboga, M. Sevkli, and S. Baskan "A Simple State-Bases Prognostic Model for Railway Turnout Systems", Proceedings of IEEE Transactions on Industrial Electronics, 2010, vol. 58, Issue 5, pp. 1718-1726.
- [13] "Intelligent Point Diagnostic System SIDIS W from Siemens Transportation Systems: Technology for Efficient Rail Services", Siemens AG Transportation Systems Rail Automation. Braunschweig, Germany, 6 p.
- [14] A.B. Nikitin, A.Yu. Panychev, and M.N. Vasilenko "A Diagnostics Method for Signal Lamp Glowers", Russian Electrical Engineering, 2016, Vol. 87, pp. 241-243, doi: 10.3103/S1068371216050126.
- [15] P. Novák, M. Daňhel, R.B. Blažek, M. Kohlík, and H. Kubátová "Predicting the Life Expectancy of Railway Fail-Safe Signaling Systems Using Dynamic Models with Censoring", IEEE International Conference on Software Quality, Reliability and Security (QRS), 25-29 July 2017, Prague, Czech Republic, pp. 329-339, doi: 10.1109/QRS.2017.43.
- [16] A. Poroshin, V. Shatokhin, A. Nikitin, and A. Kotenko "Diagnostics and Monitoring of Railway Automation and Remote Control Power Supply Devices", Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 592-597, doi: 10.1109/EWDTS.2017.8110143.
- [17] L. Heidmann "Smart Point Machines: Paving the Way for Predictive Maintenance", Signal+Draht, 2018, issue 9, pp. 70-75.
- [18] H.S. Park, H.M. Lee, H. Adeli, and I.A. Lee "New approach for health monitoring of structures: terrestrial laser scanning", Computer-Aided Civil and Infrastructure Engineering, 2007, Vol. 22, Issue 1, pp. 19-30.
- [19] Y. Park, K. Lee, C. Park, J.-K. Kim, A. Jeon, S. Kwon, and Y.H. Cho "Video Image Analysis in Accordance with Power Density of Arcing for Current Collection System in Electric Railway", The Transactions of the Korean Institute of Electrical Engineers, 2013, vol. 62, issue 9, pp. 1343-1347.
- [20] D. Efanov, G. Osadchy, D. Sedykh, D. Pristensky, and D. Barch "Monitoring System of Vibration Impacts on the Structure of Overhead Catenary of High-Speed Railway Lines", Proceedings of

- 14th IEEE East-West Design & Test Symposium (EWDTS'2016), Yerevan, Armenia, October 14-17, 2016, pp. 201-208, doi: 10.1109/EWDTS.2016.7807691.
- [21] D. Efanov, D. Sedykh, G. Osadchy, and D. Barch "Permanent Monitoring of Railway Overhead Catenary Poles Inclination", Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 163-167, doi: 10.1109/EWDTS.2017.8110142.
- [22] D. Efanov, G. Osadchy, and D. Sedykh "Protocol of Diagnostic Information Transmission via Radio Channel Concerning Health Monitoring of Infrastructure of Russian Rail Roads", Proceedings of 3rd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), St. Petersburg, Russia, May 16-19, 2017, doi: 10.1109/ICIEAM.2017.8076135.

Length Limiting of Quantum Key Distribution at Two-Stage Synchronization

Rumyantsev K.E.

*The department of information security of telecommunications system
Southern Federal University
Taganrog, Russia
rke2004@mail.ru*

Shakir H.H.Sh.

*Ministry of Education
Baghdad, Iraq
hyder.almansoor@yahoo.com*

Abstract — The synchronization subsystem is investigated for the quantum key distribution (QKD), where photon pulses are used as sync signals. The analyzed two-stage synchronization algorithm is based on the fact that the pulse-repetition period and the duration of the optical sync signals are known. Analytical expressions are obtained that establish the functional relationship between the energy parameters and probability characteristics of the single-photon synchronization subsystem, and the parameters of an optical fiber, a transmitting optical module, and a single-photon avalanche photodiode (SPAD). The requirements for the choice of the steps number in the search stage and the allowable tests number that guarantee the maximum length of the communication line are formulated. It is shown that for actually used single-photon avalanche photodiodes, the length of the fiber-optic line is limited due to the frequency of generation of dark current pulses (DCP) by a value of 51 km with a synchronization error probability of 0.05.

Keywords — *quantum key distribution, auto-compensation system, synchronization, two-stage single-photon algorithm, fiber-optic line, communication length.*

I. INTRODUCTION

The use of BB84 protocol in systems of quantum key distribution (QKD) satisfies the requirements of absolute secrecy when encrypting messages and distributing the secret key among legitimate users [1 – 3]. Thanks to the axioms of quantum physics, it ensures that hidden interception or copying of a message sent through a quantum channel is impossible. Unauthorized users can not get any information without changing the quantum state of the information carrier, which in turn will indicate the presence of an intruder in the communication line. Thus, after the next session, legitimate users can verify the presence of the intruder in the quantum communication channel.

However during tests of QKD systems (id 3110 Clavis 2 and QPN 5505) are established [4, 5], that at synchronization an average of photons in a pulse is hundreds and more. Multiphoton synchronization potentially facilitates for the intruder to organize hindrances or access to the information [6 – 8]. This procedure is based on the technical imperfection of the optoelectronic components of the QKD system and can be attributed to the Trojan horse attack [9]. The intruder becomes an integral component of the communication line between stations.

To implement this attack on the QKD system and unauthorized access into the quantum channel, the intruder must receive information about the exact time of photodetectors gating of the receiving-transmitting station [10, 11]. This is achieved by the intruder removing part of the optical power from the quantum channel during the preliminary synchronization of stations. At the synchronization stage, the control algorithms for signals

transmitted between the coding and transceiver stations do not function. Therefore, the removal of optical power from the quantum communication channel does not disrupt the operation of the QKD system and does not detect the presence of an intruder. This procedure can be implemented by using fiber-optic directional couplers or special «clothespins».

However, to increase protection against unauthorized access, it is possible to use for synchronization the photon pulses [12]. Here the photon pulse represents the optical pulse of the transmitter, attenuated to the level of registration on average less than one photon. Note that attenuation of the optical pulse to the photon level is provided when the synchronization signal propagates from the coding station to the receiving-transmitting station and is implemented by means of a controlled optical attenuator.

In [13 – 15] the analysis of one-photon synchronization algorithm of stations is carried out, in which the time frame equal to the pulse-repetition period T_s , is shared on N_w time windows with duration τ_w so, that period $T_s = N_w \tau_w$.

In the algorithm description, it is emphasized that an ideal single-photon module is an optical detector, which is capable of registering any photon. In addition, it is assumed that the photodetector does not need time to recover from the registration of a photon or a dark current pulse (DCP).

The characteristics of single-photon avalanche photodiodes (SPAD) used in QKD systems are different from an ideal single-photon photodetector. First, only one (first) photon is registered here during the SPAD analysis. Secondly, in the case of photon registration, it will take some time to restore the working state of the SPAD [16 – 18]. The total time delay between the avalanche formation and the subsequent restoration determines the insensitivity time of a single-photon photodetector. Due to the non-ideal characteristics of a real SPAD, the application of the described synchronization algorithm leads to a significant increase of the synchronization time.

In [19 – 20] the synchronization algorithm of the QKD system is proposed and investigated without dividing the time frames into time windows. The two-stage synchronization algorithm is based on the fact that at the receiving complex the pulse-repetition period T_s and the duration τ_s of optical sync signals are known.

The two-stage synchronization algorithm provides a significant time gain in short fiber-optic lines when using real SPAD compared with an algorithm in which the time frame is shared on time windows [21]. At the same time, the synchronization error probability is rather low.

However, this algorithm can be successfully applied only to short communication lines (tens of kilometers). For long fiber-optic line (about 100 km) algorithm application becomes impossible owing to considerable growth of error synchronization probability. Naturally, the restriction on the fiber-optic link length due to the synchronization is restricts the quantum key distribution length also.

Purpose. Probing a two-stage single-photon synchronization algorithm in order to formulate the conditions under which the maximum length of the communication line is guaranteed.

II. SINGLE-PHOTON ALGORITHM OF TWO-STAGE SYNCHRONIZATION

Let the synchronization start at the moment $t=0$. The equipment operates in the search (1st) stage, registering the fact of reception of a photon or DCP in the first time frame $[0, T_s]$. If in this time frame there is no excess of the amplitude discrimination threshold U_{AD} , then the search continues in subsequent intervals $[(j-1) \cdot T_s, j \cdot T_s]$, $j \geq 2$.

Let's assume, that in the moment $t_{AD} \in [(N_{step} - 1) \cdot T_s, N_{step} \cdot T_s]$, the threshold of peak discrimination is exceeded (Figure 1). The equipment goes into testing (2nd) stage, in which the re-interrogation of a single-photon photodetector $k=1, \dots, N_{test,max}$ is performed only in intervals

$$[(k-1) \cdot T_s + t_{strob1}, (k-1) \cdot T_s + t_{strob2}]. \quad (1)$$

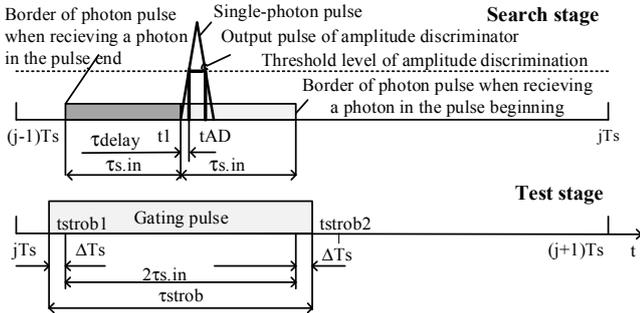


Fig. 1. A junction from a search stage to a testing stage

Note that for the rest of the time, the single-photon registration channel does not respond to the reception of photons and DCPs.

In (1) the value $t_{strob1} = t_{AD} - \tau_{delay} + T_s - 0.5 \cdot \tau_{strob}$ corresponds to the moment of the beginning of the action, and $t_{strob2} = t_{AD} - \tau_{delay} + T_s + 0.5 \cdot \tau_{strob}$ the moment of the end of the gating pulse during the re-examination. The value τ_{delay} represents the delay time between the moment of triggering the AD t_{AD} and the moment of generating a single photon t_1 .

It is established that due to a priori uncertainty regarding the time of signal reception, the duration of the gating pulse τ_{strob} should exceed the duration of the optical sync pulse more than at 2 times ¹

¹ The work does not take into account the dispersion properties of optical fiber.

$$\tau_{strob} = 2 \cdot \tau_s + 2 \cdot \Delta T_s, \quad (2)$$

where ΔT_s is the instability of the pulse-repetition period of sync pulses.

The synchronization algorithm assumes that if during the N_{test} test there was a repeated excess of the threshold level of amplitude discrimination, then a decision is made to receive a photon pulse in the analyzed time frame

$$[(N_{test}-1) \cdot T_s + t_{strob1}, (N_{test}-1) \cdot T_s + t_{strob2}]. \quad (3)$$

If, at the allowed test number $N_{test,max}$, the threshold level of amplitude discrimination has not been exceeded, the equipment returns to the search stage again.

It is necessary to dwell on the requirements to the selection of the pulse-repetition period of optical pulses T_s .

For maximum distance $L_{TF,max}$ between stations of autocompensatory system the pulse-repetition period of optical sync pulses should satisfy QKD condition

$$T_s \geq 2 \cdot L_{TF,max} / v_{OF}. \quad (4)$$

The multiplier 2 in the formula takes into account that in the auto-compensation system of the QKD a photon passes twice the fiber-optic line: *receiving and transmitting station* \rightarrow *fiber-optic line* \rightarrow *encoding station* \rightarrow *fiber-optic line* \rightarrow *a receiving and transmitting station*.

The photon propagation velocity in the optical fiber v_{OF} is determined by the refractive index of core n_{OF} and by the wavelength λ_s :

$$v_{OF} = c_{opt} / n_{OF} \quad (5)$$

where $C_{opt} = 300000$ km/s is the velocity of radiation propagation in vacuum.

III. PROBABILISTIC CHARACTERISTICS OF A TWO-STAGE SYNCHRONIZATION ALGORITHM

Poisson's law is used to describe the statistical properties of the photon flux and DCP

$$Pr\{n|\bar{n}\} = \frac{\bar{n}^n}{n!} \cdot \exp(-\bar{n}).$$

The generation probability of n events is determined by the average number of photons and/or DCP during the observation time (time frame, optical or gating pulse) \bar{n} .

Let ξ_{DCR} be the DCR generation frequency. Then, during the time frame T_s will on average be generated

$$\overline{n_{DCR.T}} = \xi_{DCR} \cdot T_s \quad (6)$$

noise pulses, and during the optical pulse τ_s

$$\overline{n_{DCR.S}} = \xi_{DCR} \cdot \tau_s \quad (7)$$

Let the first time frame be analyzed (Figure 1) during the time $[0, T_s]$. If the time $t_1 \in [0, T_s]$ corresponds to the

leading edge of an optical pulse, then two events must occur for detection. First, there should be no DCP in the interval $[0, t_1]$. Secondly, in the interval $[t_1, t_1 + \tau_s]$ at least one photon or DCP must be registered. First, confirm that you have the correct template for your paper size.

Probability of the first event

$$Pr\{n = 0 | \bar{n} = \xi_{DCR} t_1\} = \exp(-\xi_{DCR} t_1)$$

depends on the DCP generation frequency ξ_{DCR} and the random moment t_1 .

Probability of a second event

$$P_{Ds} = Pr\{n \geq 1 | \bar{n} = \overline{n_{DCR.s}} + \overline{n_s}\}$$

is determined in addition to the DCP generation frequency ξ_{DCR} and the duration of the optical sync pulse τ_s by the average number of photons during photon pulse duration $\overline{n_s}$:

$$P_{Ds} = 1 - \exp(-\overline{n_{DCR.s}} - \overline{n_s}). \quad (8)$$

The average number of registered photons (n_s) for the photon pulse duration with the fiber-optic line length is determined by the formula

$$\overline{n_s} = \overline{n_{s0}} \cdot 10^{-\frac{\alpha_{OF}[dB/km] \cdot L_{TF}[km]}{10}}, \quad (9)$$

where $\overline{n_{s0}}$ – average number of photons per pulse on the coding station exit; α_{OF} – attenuation of the optical fiber.

Thus, the conditional probability of detecting a photon pulse in the search stage during the analysis of the first frame is equal to

$$P_1\{t_1\} = \exp(-\xi_{DCR} t_1) \cdot P_{Ds}.$$

If the signal is not detected in the interval $[0, T_s]$, then it is possible in the second frame in the interval $[t_1 + T_s, t_1 + T_s + \tau_s]$. The conditional probability of detecting a photon pulse here can be calculated by the formula

$$P_1\{t_1\} = \exp(-\xi_{DCR} \cdot t_1) \cdot P_{Ds}.$$

The probability of the reception lack of photons and DCP P_{DCR0} for the time frame is determined by the average numbers of signal photons $\overline{n_s}$ and DCP $\overline{n_{DCR.T}}$ for the duration of the time frame:

$$P_{DCR0} = \exp(-\overline{n_s} - \overline{n_{DCR.T}}). \quad (10)$$

The conditional probability of detecting a photon pulse in the search stage during the j-th time frame analysis is equal to

$$P_j\{t_1\} = \exp(-\xi_{DCR} \cdot t_1) \cdot P_{DCR0}^{j-1} \cdot P_{Ds}.$$

The conditional probability of photon pulse detection in the search stage during the analysis of the first N_{step} time frames will be

$$P_D\{t_1, N_{step}\} = \sum_{j=1}^{N_{step}} P_j\{t_1\} = \exp(-\xi_{DCR} \cdot t_1) \cdot P_{Ds} \cdot \sum_{j=1}^{N_{step}} P_{DCR0}^{j-1}.$$

The series $P_1\{t_1\}, P_2\{t_1\}, \dots, P_{N_{step}}\{t_1\}$ represents a geometric progression with the denominator of the progression P_{DCR0} . Using the expression to calculate the sum of the first N_{step} members of a geometric progression, we find

$$P_D\{t_1, N_{step}\} = \exp(-\xi_{DCR} \cdot t_1) \cdot \frac{1 - P_{DCR0}^{N_{step}}}{1 - P_{DCR0}} \cdot P_{Ds}.$$

The unconditional probability (hereinafter the probability) of photon pulse detection in the search stage during the analysis of the first N_{step} frames is found by averaging the probability $P_D\{t_1, N_{step}\}$ over the probability density $\omega(t_1)$ of the photon pulse occurrence $t_1 \in [0, T_s]$

$$P_D\{N_{step}\} = \int_0^{T_s} \omega(t_1) \cdot P_D\{t_1, N_{step}\} \cdot dt_1.$$

Taking into account the equiprobable distribution of the moment of occurrence of an optical pulse $\omega(t_1) = 1/T_s$ on the interval $t_1 \in [0, T_s]$, we find

$$P_D\{N_{step}\} = \frac{1 - \exp(-\overline{n_{DCR.T}})}{\overline{n_{DCR.T}}} \cdot \frac{1 - P_{DCR0}^{N_{step}}}{1 - P_{DCR0}} \cdot P_{Ds}. \quad (11)$$

Assuming the possibility of an infinite number of steps in the analysis of frames $N_{step} \rightarrow \infty$, the maximum detection probability of a photon pulse in the search stage is found

$$P_{D.max} = \frac{1 - \exp(-\overline{n_{DCR.T}})}{\overline{n_{DCR.T}}} \cdot \frac{1}{1 - P_{DCR0}} \cdot P_{Ds}. \quad (12)$$

The factor $1 - P_{DCR0}^{N_{step}}$ in (11) determines the allowable deterioration of the detection probability ΔP_D in the search stage with the number of steps limited to N_{step} . This allows us to formulate requirements for the choice of the minimum allowable number of steps in the search stage:

$$N_{step} \geq \frac{\ln(\Delta P_D)}{\ln(P_{DCR0})} \quad (13)$$

Let during the time τ_{strob} in the time interval (3) the average number of recorded photons and DCPs is

$$\overline{n_{strob}} = \overline{n_s} + \xi_{DCR} \cdot \tau_{strob}. \quad (14)$$

Then the probability of an error about the detection of an optical sync signal in the testing stage (the absence of photon registration and DCP for the entire testing time) can be calculated by the formula

$$P_{err.test} = \exp(-N_{test.max} \cdot \overline{n_{strob}}). \quad (15)$$

The expression allows to formulate the requirements for the selection of an admissible number of tests $N_{test.max}$ to ensure a given error probability $P_{err.test0}$ in making a decision about the detection of an optical synchronization signal in the testing stage:

$$N_{test.max} \geq \frac{1}{\overline{n_{strob}}} \cdot \ln\left(\frac{1}{P_{err.test0}}\right). \quad (16)$$

For the described algorithm, the probability of an error according to the results of two stages of synchronization, taking into account (11) and (15), will be

$$P_{err.sync} = 1 - P_D\{N_{step}\} \cdot (1 - P_{err.test}) \quad (17)$$

From (15) it can be seen that increasing the number of tests can make a small probability of an error in the decision to detect an optical sync signal in the testing stage $P_{err.test}$. However, to achieve a close to 1 the detection probability in the search stage, increasing the analysis time in the search stage is impossible. Indeed, with $\overline{n_{DCR.T}} \ll 1$, $\overline{n_s} < 0.1$ the formula (12), taking into account (8) and (10), is converted to

$$P_{D.max} = \frac{\overline{n_s}}{\overline{n_{DCR.T}} + \overline{n_s}} = \frac{1}{1 + \overline{n_{DCR.T}}/\overline{n_s}}. \quad (18)$$

It can be seen that for the small fiber-optic line length the condition $\overline{n_{DCR.T}} \ll \overline{n_s}$ is satisfied. As a result, with $P_{err.test} = 1$, we have $P_{err.test} = 1$ and $P_{err.sync} \cong 1$. However, already at $\overline{n_{DCR.T}}/\overline{n_s} = 0.1$ we find that the probability of a synchronization error exceeds 10 %, at 0.2 - 17 %, and 0.5 - already 33 %.

In order for the synchronization error probability not to exceed the allowable value $P_{err.sync.lim}$, the condition

$$\overline{n_s} \geq \overline{n_{DCR.T}} \cdot \frac{1 - P_{err.sync.lim}}{P_{err.sync.lim}}. \quad (19)$$

The last condition determines the maximum length of the fiber-optic line $L_{TF.max}$. Indeed, in view of (9), condition (19) is converted to

$$\overline{n_{s0}} \cdot 10^{-\frac{\alpha_{OF}[dB/km] \cdot L_{TF.max}[km]}{10}} \geq \overline{n_{DCR.T}} \cdot \frac{1 - P_{err.sync.lim}}{P_{err.sync.lim}}.$$

Note that with increasing the fiber-optic line length, according to (4), the pulse-repetition period of optical sync pulses T_s increases and, as a consequence, the average number of DCP per time frame $\overline{n_{DCR.T}}$ increases. On the other hand, according to (9), the average number of registered photons $\overline{n_s}$ decreases.

According to (4), (6) and (9), the maximum fiber-optic line length can be found by solving the transcendental equation

$$\frac{P_{err.sync.lim}}{1 - P_{err.sync.lim}} 10^{-\frac{\alpha_{OF}[dB/km] \cdot L_{TF.max}[km]}{10}} = \frac{\overline{n_{s0}} \cdot v_{OF}[km/s]}{2 \cdot \xi_{DCR}}. \quad (20)$$

From the formula it can be seen that the optic fiber length is maximum when using an optical fiber with zero attenuation

$$L_{TF.lim}[km] = \frac{\overline{n_{s0}} \cdot v_{OF}[km/s]}{2 \cdot \xi_{DCR}} \cdot \frac{P_{err.sync.lim}}{1 - P_{err.sync.lim}}. \quad (21)$$

Figure 2 shows the dependences of the maximum fiber-optic line length on the DCP generation frequency, using formulas (1) - (21) with the following initial data: single-mode optical fiber Corning SMF-28e (attenuation 0.20

dB/km; refractive $\overline{n_{s0}} = 0.1$); given probability of synchronization error $P_{err.sync.lim} = 0.05$. The error in calculating the length by (20) did not exceed 0.1 %.

It can be seen from the figure that the fiber-optic line length with a frequency of DCP generation of 10 Hz does not exceed 20.7 km. This is more than 2 times less than the maximum fiber-optic line length with the exclusion of attenuation in the optical fiber (53.8 km). The pulse-repetition period is 200 ms.

Table 1 shows the characteristics of photodetector modules based on SPAD used in QKD systems. The last column shows the results of calculations of the maximum line length on an optical fiber with an attenuation of 0.20 dB/km, which can be achieved using the module for receiving optical radiation with a wavelength of 1550 nm and guaranteeing the probability of synchronization error not worse than 0.05. The parameter values for the photodiode module id210 are given at a quantum efficiency of the photocathode of 10 %, and in brackets at 20 %.

The analysis shows that even with the use of SPAD with a minimum DCP generation frequency of 0.4 Hz, the maximum line length does not exceed 70 km. Moreover, when more stringent requirements are imposed on the synchronization efficiency, the QKD length is reduced. For example, when the permissible value of the synchronization error probability changes from 0.05 to 0.01, the maximum fiber-optic line length drops 2.8 times from 20.7 to 7.4 km at a DCP generation frequency of 10 Hz.

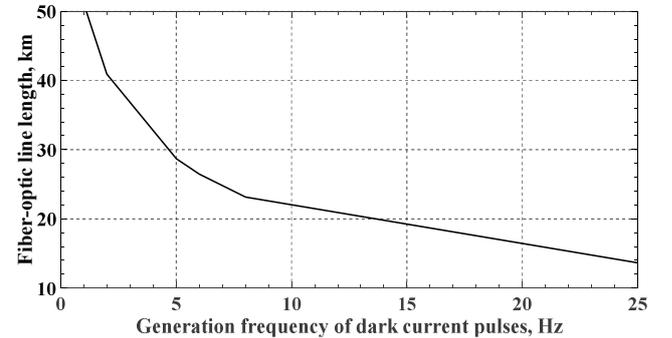


Fig. 2. Dependence of the maximum length of fiber-optic line from the DCP generation frequency

TABLE 1. PARAMETERS OF PHOTORECEIVER MODULES BASED ON SPAD

Photoreceiver module name	The DCP frequency, not more, Hz	The maximum fiber-optic line length, km
id201	100	4.4
id210-SMF-A. Ultra-Ultra Low Noise	0.4 (2)	65.6 (40.9)
id210-SMF-B. Ultra-Low Noise	1 (2)	51.1 (40.9)
id210-SMF-C. Standard	6 (30)	26.5 (10.9)
id210-MMF	8 (40)	23.1 (8.9)
id230. Ultra-Low Noise	25	12.2
id230 Standard	50	7.6
id280	100	4.4

IV. ACKNOWLEDGMENT

The article was prepared with the support of the Russian Foundation for Basic Research, project 16-08-00752.

V. CONCLUSION

Analytical expressions are obtained that establish the functional relationship between the energy parameters and probability characteristics of the single-photon synchronization subsystem, and the parameters of an optical fiber, a transmitting optical module, and a single-photon avalanche photodiode.

Requirements for the choice of the minimum number of steps in the search stage and the allowable number of tests in the testing stage, which guarantee the maximum length of the communication line, are stated. It has been confirmed that by increasing the number of tests, it is possible to make an arbitrarily small probability of an error in making a decision on the detection of an optical sync signal in test mode. However, to achieve an arbitrarily close to 1 probability of detection of a photon pulse in the search stage, increasing the analysis time in the search stage is impossible.

It is shown, that the maximum fiber-optic link length is, above which it is impossible to construct system QKD, even using an optical fiber with zero attenuation. For used single-photon avalanche photodiodes, the fiber-optic line length is limited due to the DCP generation frequency of 70 km with a synchronization error probability of 0.05.

REFERENCES

- [1] Bennett C., Brassard G. Quantum cryptography: Public key distribution and coin tossing // Proceedings of IEEE international conference on computers, systems and signal processing. Bangalore, India. – New York: Institute of Electrical and Electronics Engineers, 1984. – P. 175-179.
- [2] Gisin N., Ribordy G., Tittel W., Zbinden H. Quantum cryptography // Reviews of Modern Physics. – 2002. – Vol. 74. – № 1. – P. 145-195.
- [3] Shor P.W., Preskill J. Simple proof of security of the BB84 quantum key distribution protocol // Physical Review Letters. 2000. Vol. 85. P. 441-444. Quant-ph/0003004.
- [4] Pljonkin A., Rumjantsev K. Preliminary stage synchronization algorithm of auto-compensation quantum key distribution system with an unauthorized access security // Proceeding of the 15th International Conference on Electronics, Information, and Communication 2016 (ICEIC 2016). Jan 27-30, 2016. Danang, Vietnam. DOI: 10.1109/ELINFOCOM.2016.7562955.
- [5] Kurochkin V.L. and other. Experimental studies in the field of quantum cryptography // Photonics. - 2012. - V. 5. - P. 54-66.
- [6] Gerhardt I., Liu Q., Lamas-Linares A., Skaar J., Kurtsiefer C., and Makarov V. Full-field implementation of a perfect eavesdropper on a quantum cryptography system. Nat. Commun. 2, 349 (2011).
- [7] Hacking commercial quantum cryptography systems by tailored bright illumination / Lars Lydersen, Carlos Wiechers, Christoffer Wittmann, Dominique Elser, Johannes Skaar, Vadim Makarov // Nature Photonics. 29 August 2010. DOI:10.1038/nphoton.2010.214.
- [8] Makarov V. Controlling passively quenched single photon detectors by bright light // New Journal of Physics. 2009. Vol. 11. 065003.
- [9] Gisin N., Fasel S., Kraus B., Zbinden H., Ribordy G. Trojan-horse attacks on quantum-key-distribution systems // Physical Review A. 2006. Vol. 73. 022320.
- [10] Rumyantsev K.E. Synchronization in the system of quantum key distribution with automatic compensation of polarization distortions // Telecommunications. 2017. № 2. P. 32 - 40.
- [11] Rumyantsev K.E. Protection of the synchronization process in the system of quantum key distribution with automatic compensation of polarization distortion // Telecommunications. 2017. №3. P. 36-44.
- [12] Rumyantsev K.E., Plyonkin A.P. Synchronization of the quantum key distribution system in the single-photon registration mode of pulses to increase security // Radio and communications technology. 2015. № 2. P. 125-134.
- [13] Rumyantsev K.E., Plyonkin A.P. Improving the efficiency of the algorithm for entering into synchronism of the system of quantum key distribution. Izvestiya SFedU. Engineering sciences. 2015. No. 8 (169). P. 6-19.
- [14] Pljonkin A., Rumjantsev K. Synchronization algorithm of quantum key distribution system with protection from unauthorized access // Proceeding of the IEEE Photonics Society Workshop on Recent Advances in Photonics (IEEE WRAP 2015). December 16 -17, 2015. Bangalore, India, KA. 7805988.
- [15] Pljonkin A., Rumjantsev K. Single-photon Synchronization Mode of Quantum Key Distribution System // Proceeding of the International Conference on Computational Techniques in Information and Communication Technology 2016. (ICCTICT 2016). March 11-13, 2016. New Delhi, India. P.531-534. 7514637.
- [16] ID100 v 2016 01 28. Specifications. 2016. www.idquantique.com/
- [17] ID230 v2015 04 29. Specifications as of May 2015.
- [18] ID280. <http://www.idquantique.com/photon-counting/photon-counting-modules/id280/>
- [19] Rumyantsev K., Rudinsky E. Parameters of the two-stage synchronization algorithm for the quantum key distribution system // Proceedings of the 10th International Conference on Security of Information and Networks (SIN-2017). 13-15.10.2017. Rajasthan, India. DOI 10.1145/3136825.3136888. P. 140-150.
- [20] Rumyantsev K., Rudinsky E. Time synchronization method in quantum key distribution system with automatic compensation of polarization distortions // Proceedings of the 2nd International Conference on Multimedia and Image Processing, ICMIP 2017. Wuhan, China, March 17-19, 2017. P. 346-349. 132083. DOI: 10.1109/ICMIP.2017.68/
- [21] Rumyantsev K.E., Rudinsky E.A. Two-stage time synchronization algorithm in the system of quantum key distribution with automatic polarization distortion compensation // Izvestiya SFU. Technical science. 2017. № 5. P. 75-89.

Technological Foundations of Traffic Controller Data Support Automation

Joseph M. Kokurin,
DSc, Solomenko Institute of transport problems of the Russian
Academy of sciences, Chief Scientific Problem Lab Employee
Transport Organizations Systems, Professor,
Saint Petersburg, Russia,
kokyrinim@mail.ru

Dmitrii V. Efanov,
DSc, Professor at "Automation, Remote Control
and Communication on Railway Transport",
Russian University of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

Abstract—The paper provides rationale for the technological and technical foundations of data support automation for decision-making by a traffic controller, who controls train traffic and local performance in their subdivision. The traffic controller solves the most complex tasks in case of significant out-of-schedule train traffic unexpectedly caused by delays, changes in governed speed and other reasons. In such a case, the controller needs to change train overtaking stations specified in the tight run profile, predict random time of train arrival at stations of the subdivision relying on the knowledge of local conditions and experience. Speed-time-distance calculations are expected to improve the accuracy of train performance prediction, which will allow overtake stations to be identified based on the criteria of reduced stops of overtaken trains and avoided delays of overtaking ones. Established regularities and irregularities serve as a basis for creating an automated train management system, which constitutes the middle level of the hierarchy of rail traffic process control systems.

Keywords: railway sector; traffic planning and control; traffic controller; data support automation for decision-making; train stop and overtaking time minimization.

I. INTRODUCTION

The key goal of the railway sector is timely transportation of passengers and cargo to their destinations. To do this, the railway sector combines infrastructure components, including power supply facilities, as well as rolling stock and train traffic control and management systems [1–3]. The sector is divided into railways, control areas and subdivisions controlled by traffic controllers. By processing real-time data, they handle trains according to the tight or untight run profile.

Complexity of the traffic process, which enhances during regular maintenance, upgrade and repair of infrastructure, dictates only partial automation of controller's decision-making and execution processes. This issue is being studied both in Russia and abroad. The authors of this paper focus on the traffic controller data support automation.

II. TRAFFIC CONTROLLER DATA SUPPORT AUTOMATION PHILOSOPHY

The main resource in railway traffic planning is the time of day intervals, which are specified in the tight run profile for each train to pass through the subdivision. The key benefit of train traffic scheduling is that controllers and all involved parties have the most correct prediction of train arrival and departure times for all stations of the subdivision. This provides all transportation personnel, and, above all, traffic controllers with the best data support to restore train

traffic on schedule in case of random delays. The complex task in case of significant delays is to change overtake stations specified in the tight run profile. In such a case, the controller has to predict specified points in time of overtaken and overtaking trains based on real-time statistics on the time spent by each train involved in overtaking to pass a section of the subdivision.

The solution of this complex task cannot be automated based on the queuing theory, as this theory considers train traffic without predicting the specified points in time. It is therefore proposed to use simulation based on speed-time-distance calculations [4] of real trains with known properties (weight, length, traction performance of the loco, running resistance, track layout and grading).

Studies show that traffic controller data support automation should reliably ensure timely receipt of all adequately accurate data and sufficient time for making, adjusting and executing optimal decisions. The criterion for this is minimization of train stop and movement time.

It is proposed to consider the automated decision-making and execution technique when selecting a station not specified in the tight run profile for overtaking of train $j-1$ by train j by building their time travel lines $t_{x,j-1} = f(S)$ and $t_{x,j} = f(S)$ determined by speed-time-distance calculations depending on distance S traveled by the center of gravity of each train (Fig. 1).

Such an approach accurately and visually links train travel times to the distance position, which is necessary to determine signal aspects of wayside and on-board light signals correlated to the train position. The computational solution of the nonlinear differential equation of train motion [5] should be made with a sufficiently small integration step for distance ΔS , which ensures the required accuracy of calculations. Positions shall be increased by half the length of the train when determining the times of infrastructure occupancy by train and decreased when determining the times of clearing.

To derive the train acceleration function $a = \frac{dv}{dt}$ of the specific resultant force $(f_k \pm \omega_k - b_t)$, we use the Newton's second law:

$$S = \sum \Delta S; \quad (1)$$

where $(f_k \pm \omega_k - b_t)$ is locomotive forces acting on a train, running resistance and braking depending on the train speed v , and ζ is the acceleration of train forward motion factoring in the rotation of rolling stock parts.

Integration of the resultant differential equation produces functions of speed v , time t and distance traveled S . By integrating within the time change from t_n to t_k and speed from v_n to v_k , we obtain:

$$\Delta t = t_k - t_n = \frac{1}{\zeta} \int_{v_n}^{v_k} \frac{dv}{f_k \pm \omega_k - b_t}. \quad (2)$$

By replacing in equation (1) dt with $\frac{ds}{v}$ and integrating both parts, we find the distance: $\Delta S = S_k - S_n$ traveled by train with the given speed variation:

$$dS = \frac{v dv}{\zeta(f_k \pm \omega_k - b_t)}; \quad \Delta S = \frac{1}{\zeta} \int_{v_n}^{v_k} \frac{v dv}{f_k \pm \omega_k - b_t}. \quad (3)$$

Within small speed delta $\Delta v = v_k - v_n$, the resultant force can be considered constant and equal to the average. By integrating the right parts of the equations, we obtain:

$$\Delta t = \frac{v_k - v_n}{2\zeta(f_k \pm \omega_k - b_t)}; \quad \Delta S = \frac{v_k^2 - v_n^2}{2\zeta(f_k \pm \omega_k - b_t)}. \quad (4)$$

When motive forces equal resistance forces, the resultant force becomes zero. At this point in time, train speed shall be assumed equal to that achieved until the resultant force appears.

By replacing $v_k^2 - v_n^2$ with $(v_k + v_n)(v_k - v_n)$ and using $\frac{v_k + v_n}{2}$ as the average speed, we obtain: $\Delta S = \Delta t \frac{v_k + v_n}{2}$.

To determine the estimated ratios with integration step ΔS , by replacing v_k with $v(S + \Delta S)$, and v_n with $v(S)$, we obtain:

$$S = \sum \Delta S; \quad t(S + \Delta S) = t(S) + \frac{2\Delta S}{v(S + \Delta S) - v(S)}. \quad (5)$$

For priority train j to overtake train $j-1$, station i may be used provided that the predicted distance between

these trains $L_{j-1,j}$, taking into account speed reduction of the overtaken train to stop on the side track of the overtake station, is sufficient to prevent a yellow aspect on the on-board light signal of the overtaking train that requires speed reduction. Speed reduction is required, as a rule, when passing a wayside light signal with a yellow aspect, point switches and approaching a light signal with a red aspect. The specified train-to-train distance depends on the number of signal aspects of automatic block signaling and cab signaling, train-to-train interval, the difference in train speeds and block lengths, and the availability of overlaps.

Railway automation systems currently employ the principles of light signaling using wayside and on-board light signals to transmit data about train speeds [6–8]. By the way, EU railways already intend to abandon color light signaling and switch to digital technologies. The Russian railways are not currently in active discussion of this transition and intend to keep conventional data transmission to the driver. For this purpose, the following main signal aspects are used to avoid collision: accepted terminology: “green light” is clearance to move with governed speed when two or more blocks are vacant with three-aspect light signaling; “yellow light” is clearance to move with reduced speed when one block is vacant, ready to stop before the next light signal; “red light” is prohibition to pass the signal. Four-aspect light signaling supplemented by signal aspect “simultaneous yellow and green lights” is used in blocks with heavy suburban traffic.

Signal aspects “two simultaneous yellow lights” and “two simultaneous yellow lights with upper one blinking” and other signal aspects are used to reduce speed when moving with point switch deflection at the entrance signal that clears or prohibits movement to the station.

Cab signaling transmits signal aspects to the locomotive in front of the wayside light signal when the train occupies the track circuit (block) before this light signal [9]. Codes in the form of pulse trains are transmitted to track circuits: “G” (“green” – three pulses in the transmission cycle), “Y” (“yellow” – two pulses with a short pause between them) and “R-Y” (“red-yellow” – two pulses with a long pause between them). Receipt and decoding of codes by on-board equipment provide a corresponding signal aspect on the on-board light signal.

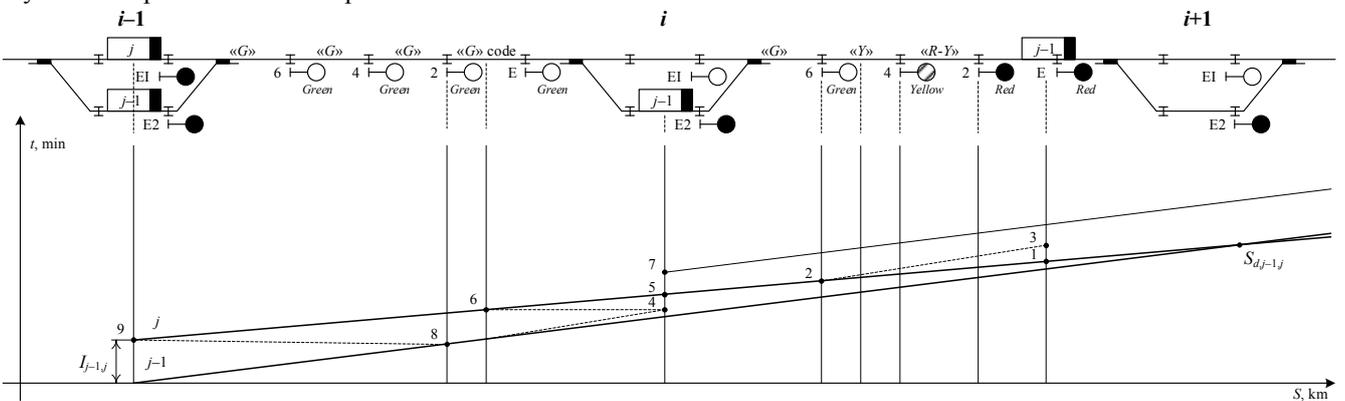


Fig. 1. Overtake station selection technique.

The overtake station selection should begin with determining the distance point $S_{d,j-1}$, where the predicted lines of overtaken $j-1$ and overtaking j trains intersect. When the overtaken train reaches this point, the train separation system (automatic block signaling, cab signaling or other) will reduce the speed of the overtaking train to avoid collision of trains.

In one of the possible predictable scenarios shown in Fig. 1, the overtaken train $j-1$ is cleared to the station side track $i+1$ with a stop for non-stop pass of the overtaking train, with signal aspect “two yellow lights” on E light signal. Code “ Y ” is transmitted from this light signal to the track circuit to light signal 2. When the overtaken train enters this track circuit, a yellow aspect lights up on the on-board light signal. The driver or the automatic train operation system will reduce the train speed along the dashed line, which intersects with the traveling line of the overtaking train at point 1 shifting the point $S_{d,j-1}$ to this point.

At the predicted time that is point 1, the overtaken train will occupy the block behind light signal 2. This light signal will give “red light”, while light signal 4 will give “yellow”, from which code “ Y ” will be sent to the block to light signal 6. When passing light signal 6, the driver of the overtaking train will see a yellow aspect on the on-board light signal and will begin to slow down (line 2–3). This will lead to a delay of the overtaken train and the loss of energy for traction.

Therefore, station $i+1$ is not suitable for overtaking the train $j-1$ and it is necessary to check the possibility of overtaking this train at station i .

Signal aspect “two yellow lights” is used at light signal E to clear the overtaken train to the second track of station i . Code “ Y ” will be sent to the block between light signals E and 2, and light signal 2 will turn on “yellow blinker light”. Code “ G ” will be sent to the block between light signals 2 and 4, and light signal 4 will turn on “green light”. When passing light signal 2, the driver of the overtaken train will begin to slow down, light signal 2 will turn on “red light”. Line 8–9 will determine the overtaking train position at station $i-1$ at this time in front of light signal E with “green light”.

At the time of arrival of the overtaken train $j-1$ at the overtake station i , taking into account the deceleration caused by the stop at the overtake station (line 8–4), the overtaking train (line 4–6) will be in front of light signal E with “green light”.

Consequently, the overtaking train will not be delayed when overtaking at station i .

The interlocking system records times of overtaken train $t_{j-1,r,i}$ clearance and isolated section release as the time of train arrival at the station allowing the entrance signal to be opened to the overtaking train. Slowdown for the time $\tau_{j-1,dl,i}$ of the isolated section release after clearance of train $j-1$ is made to prevent point switching under the train dur-

ing a short-term loss of the train shunt by the overtaken train. To open the entrance signal to the overtaking train j , the station attendant or traffic controller needs to spend time $\tau_{j,li}$ on decision-making, sending orders to prepare the route by using the control panel, switching points on the route and opening the entrance signal. All these times shall be included in the predicted time $t_{p,j-1,i}$ of arrival of the overtaken train $j-1$ to station i of the planned overtake:

$$t_{p,j-1,i} = t_{j-1,r,i} + t_{j-1,dl,i} + t_{j,li}. \quad (6)$$

Time $t_{p,j-1,i}$ thus corresponds to the time of entrance signal opening to train j for non-stop pass via main-line track 1T of station i , when the overtaken train $j-1$ is located within the useful length of side station track 2T.

The time interval (see Fig. 1) formed at the intersection of the center of gravity line of the overtaken train at station i with the train travel lines (points 4 and 5) is equal to the predicted interval of concurrent arrival of trains $j-1$ and j at the overtake station. These intervals shall be used to make a predictable run profile.

When overtaking at station $i-1$, the overtaken train will arrive at this station earlier than at station i , and its stop pending the passage of the overtaking train will increase [10].

Therefore, station i should be used for overtaking if there are no additional local restrictions (presence of a station track of the appropriate useful length, the possibility of stopping the overtaken heavy train, etc.).

With a small difference in train speeds, their travel lines may not intersect within the subdivision. In this case, the traffic controller will need to identify the overtake station using speed-time-distance calculations to obtain the exact times of arrival and departure of trains at stations of the block considering the location of trains, conditions of their movement and local features of the block.

The above description of the overtake station selection technique is the basis for developing an algorithm for automatic solution of this problem. However, prioritization of train j over train $j-1$ should be left to the traffic controller at the first stage of automation given the complexity of this process.

III. OVERTAKE STATION SELECTION ALGORITHMS

The following set of rules is proposed to develop an automatic overtake station selection algorithm.

Rule 1. If the overtaking train catches up with the overtaken one on the haul, then overtake at the station in front of it will cause a delay of the overtaking train. If the delay is unauthorized, the possibility of overtake at the previous station should be checked.

Rule 2. If the overtaken train timely clears the entrance route for the overtaking train, which will open the entrance signal at a time that does not cause unacceptable speed re-

duction of the overtaking train, then the station in question should be used for overtake.

Rule 3. The transfer of overtake to the previous station $i-1$ will increase stopping time of the overtaken train pending the overtaking train.

$$(B)t_{p_{j-1}}\left(S_{j-1} + \sum_{i=1}^m \Delta S_i\right) = t_{p_j}\left(S_j + \sum_{i=1}^m \Delta S_i\right) + I_{j-1,j}(A) \begin{cases} \langle\langle \text{No} \rangle\rangle & S_m = S_j + \sum_{i=1}^m \Delta S_i > S_k \\ \langle\langle \text{Yes} \rangle\rangle & S_m = S_j + \sum_{i=1}^m \Delta S_i \end{cases} + \Delta S_i B \quad (7)$$

In formula (7), A is the algorithm operator that checks for the equality of predicted times, the movement of both trains, which corresponds to the location of their centers of gravity in the same position $S_{d_{j,j-1}}$, m is the number of integration and summation steps $+\Delta S_i$, when the overtaking train catches up with the overtaken one. S_j and S_{j-1} are positions of the overtaking and overtaken trains at the moment of running time start.

If the equality is not achieved (the comparison result is “No”), then the distance traveled is increased by $+\Delta S_i$, the calculation of the new result is the comparison of running times of both trains.

If the distance traveled by the overtaking train $S_m = S_j + \sum_{i=1}^m \Delta S_i > S_k$ exceeds the subdivision end point S_k , the controller shall select the overtake station.

The algorithm for determining the distance point, where the overtaking train will catch up with the overtaken one, can be represented as follows:

If the equality of times is achieved (the comparison result is “Yes”), then the overtaking train position $S_m = S_j + \sum_{i=1}^m \Delta S_i$ is calculated and determined relative to the station in front.

In the case under consideration (see Fig. 1), the overtaking point is located in the block in front of station $i+1$ between light signals 4 and 6, which will delay the overtaking train and mean that this station is unsuitable for overtake.

The possibility of using station i for overtaking the train $j-1$ is determined at the time of entrance signal opening at station i to clear the overtaking train when the overtaken train stops on a side track:

$$(B)t_{p_j}\left(S_j + \sum_{i=1}^n \Delta S_i\right) + I_{j-1,j} = t_{p_{j-1}}\left(S_{j-1} + \sum_{i=1}^n \Delta S_i\right) + \tau_{p_{j-1,j}}(A) \begin{cases} \langle\langle \text{No} \rangle\rangle & S_j + \sum_{i=1}^n \Delta S_i \\ \langle\langle \text{Yes} \rangle\rangle & \end{cases} + \Delta S_i B$$

$$\left(S_{j-1} + \sum_{i=1}^n \Delta S_i\right) - \left(S_j + \sum_{i=1}^n \Delta S_i\right) \geq N_{bl}^a(A) \begin{cases} \langle\langle \text{No} \rangle\rangle & \text{Overtake station } i-1 \\ \langle\langle \text{Yes} \rangle\rangle & \text{Overtake station } i \end{cases} \quad (8)$$

where B is the algorithm transition to operator (B) ; n is the number of calculation steps, at which the overtaken train clears route sections for the overtaking one; $\tau_{p_{j-1,j}}$ is the time spent on slowing down the route section release for the overtaking train and entrance signal opening; N_{bl}^a is the number of blocks between the overtaken and overtaking trains allowing the overtaking train to move with on-board green signal aspects (the upper index indicates the number

of color light signal aspects, with three-aspect color light signaling $a=3$ and $N_{bl}^a=2$).

If the condition being checked is not met, the overtaking train will move with the yellow on-board light signal and the train $j-1$ should be overtaken at the station $i-1$. If the condition is met, the distance between trains allows movement of the overtaking train without delay with green aspects of the on-board light signal, and overtake should be performed at the station i .

The overtake station selection solution recommended by the automated system is presented to the traffic controller in the accepted form of a predictable run profile (string-line diagram) for the feasibility analysis [11 – 13]. Higher level of decision-making automation with confirmation by the driver of the selected “scenario” is possible in the future as data transmission reliability, accuracy and security technologies become more advanced [14].

IV. CONCLUSION

Overtake station selection algorithms described herein may be adapted to the software of train dispatching systems at subdivisions. This will allow a system to be implemented for train performance computation directly at the traffic controller’s workplace (also transmit these calculation data automatically to on-board automation equipment as technology further advances).

The use of speed-time-distance calculation software at the traffic controller’s workplace enhances data support automation, including train arrival and departure prediction, and optimal selection of overtake stations in case of deviations from the tight run profile.

REFERENCES

- [1] D. Efanov, and G. Osadchy “Paradigms for Building Control Systems on Railroad Transport: from the Systems of Electrical Interlocking of Points and Light Signals to Smart Grid Train Movements Controlling Systems”, Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS’2018), Kazan, Russia, September 14-17, 2018, pp. 213-220, doi: 10.1109/EWDTS.2018.8524809.
- [2] L. Arend, L. Pott, N. Hoffmann, and R. Schanck “ETCS Level 2 without GSM-R”, Signal+Draht, 2018, (110), 10, pp. 18-28.
- [3] F. Prüter, and P. Hintze “But That’s Not the Kilometre in the Plan!” – The Potential of Georeferenced Railway Infrastructure Data”, Signal+Draht, 2018, (110), 11, pp. 6-15.
- [4] “Rules of Traction Calculations for Train Work” (in Russ.), Approved by the order of Russian Railways of May 12th, 2016, No. 867 r.
- [5] I.M. Kokurin, and L.F. Kondratenko “Operational Basics of Railway Automation and Remote Control Devices” (in Russ.), Moscow, Transport, 1989, 184 p.
- [6] T. Takashige “Signalling Systems for Safe Railway Transport”, Japan Railway & Transport Review 21, 1999, pp. 44-50.
- [7] C. Hall “Modern Signalling: 5th edition”, UK, Shepperton: Ian Allan Ltd, 2016, 144 p.
- [8] G. Theeg, and S. Vlasenko “Railway Signalling & Interlocking: 2nd Edition”, Germany, Hamburg: PMC Media House GmbH, 2018, 458 p.
- [9] V. Shamanov “Formation of Interference from Power Circuits to Apparatus of Automation and Remote Control”, Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS’2018), Kazan, Russia, September 14-17, 2018, pp. 140-146, doi: 10.1109/EWDTS.2018.8524676.
- [10] I.M. Kokurin, and A.B. Vasiliev “Automation of Decision-Making Information Support for Train Dispatcher for Train Traffic Organization” (in Russ.), Automation on Transport, 2015, Vol. 1, Issue 2, pp. 156-167.
- [11] I.M. Kokurin “Theoretical and Technological Foundation of Constructing a Self-Organizing Centralized Traffic Control System” (in Russ.), Automation on Transport, 2017, Vol. 3, Issue 3, pp. 345-354.
- [12] M. Reakes “Management of Integrated Training Systems”, IEEE Conference on Aerospace and Electronics, Dayton, OH, USA, 21-25 May 1990, vol. 2, pp. 924-928, doi: 10.1109/NAECON.1990.112894.
- [13] D. Pan, Y. Zheng, and C. Zhang “On Intelligent Automatic Train Control of Railway Moving Automatic Block Systems Based on Multi-Agent Systems”, Proceedings of the 29th Chinese Control Conference, Beijing, China, 29-31 July 2010, pp. 4471-4476.
- [14] D.V. Gavzov, V.V. Sapozhnikov, and V.I. Sapozhnikov “Methods for Providing Safety in Discrete Systems”, Automation and Remote Control, 1994, vol. 55, issue 8, pp. 1085-1122.

Sum Codes of Weighted Data Bits for Objectives of Automation Logical Devices Technical Diagnostics

Dmitry Efanov,
DSc, Professor at "Automation, Remote
Control and Communication
on Railway Transport",
Russian University of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

Valery Sapozhnikov,
DSc, professor at "Automation and Remote
Control on Railways" Department,
Emperor Alexander I St. Petersburg State
Transport University,
St. Petersburg, Russia
port.at.pgups@gmail.com

Vladimir Sapozhnikov,
DSc, professor at "Automation and Remote
Control on Railways" Department,
Emperor Alexander I St. Petersburg State
Transport University,
St. Petersburg, Russia
at.pgups@gmail.com

German Osadchy,
Technical Director of Scientific and Technical Center
"Integrated Monitoring Systems" LLC,
St. Petersburg, Russia
osgerman@mail.ru

Teng Teng,
PhD student of "Electrical Engineering" Faculty,
Jiangsu Normal University,
Jiangsu, China
liliva.ten@list.ru

Abstract—The properties of codes with the summation of weighted data bits regarding detection of errors in the data vectors are researched by the authors. Such a problem arises if the sum code is used as some basis for the technical diagnostics system, for example, the concurrent error-detection system of logical devices of automation and computing hardware. Characteristics of error detection ability determine the synthesis strategy of technical diagnostics equipment in the fault detection systems via one or another sum code. A brief review of the use of classical and weighted sum codes for the purpose of discrete systems technical diagnostics problem solving is given in the article. Classification of sum codes is presented. The error detection characteristics in data vectors via weighted sum codes according to types (unidirectional, symmetrical and asymmetrical ones) and multiplicities are analyzed in detail. It has been proved that weighted sum codes, which do not use the deduction operation on a predetermined module during plotting, have the property of identifying any unidirectional distortions in the data vectors. Besides, it has been proved that weighted sum codes cannot have a uniform distribution of data vectors between all check vectors, which means that it is impossible to plot a weighted sum code with a theoretical minimum of the total number of undetectable errors. A method for implementing weighted code generators with summation, based on the use of standard adders and half adders' circuits, is presented. This approach is translated to the use of multiplexers being the part of the structures of modern programmable logic integrated circuits and widely used for development of modern automation systems and computing equipment.

Keywords—technical diagnostics, discrete systems, concurrent error-detection systems, sum codes, Berger code, weighted sum code

I. INTRODUCTION

Technical diagnostics is one of the most important procedures used to specify the technical condition of the units of automatic control systems per objects both in transport and in industry [1 – 3]. Generally, this procedure is fulfilled in two modes: 1) during disconnection of the object under diagnostics from the controlled objects (or in a specially

assigned time period prior to using the object under diagnostics in computational processes) as well as during application of special inspection actions to its inputs (test diagnostics); 2) without disconnecting the object under diagnostics from the controlled objects when the working actions are diagnostic ones (working, or functional, diagnostics) at the same time. In this case, the procedure for applying a fault detection test is fulfilled as a rule, which allows specifying whether the object under diagnostics is operative or inoperative. In case of recording malfunctions in the operation of the object under diagnostics, the diagnostic test is applied to its inputs already, which makes it possible to localize the defect.

Development of technical diagnostics equipment synthesis methods, as well as test synthesis methods for blocks and components of diagnostic systems is one of the most important tasks of technical diagnostics. The versatile approach based on duplication (and even triple modular redundancy) of the object under diagnostics and comparison of the calculation data is known and well represented [4 – 7]. Application of this approach makes it possible to identify any faults occurring in the object under diagnostics, however, it requires big expenses for system implementation, as well as high overhead costs during its operation (for example, expenses for power consumption and heat removal arrangement). Identification of the whole variety of failures is not required in many applications and their variety is limited (moreover, many of them are covered by simpler fault models, for example, via stuck-at fault models). For example, the ability to identify 100% of single stuck-at faults is one of the basic requirements regarding reliability and safety for railway automation and remote control devices [8]. Focusing on a specific fault model and limiting their set, it is possible to synthesize diagnostic systems, in which technical diagnostics equipment (supervision circuits) will have a reduced implementation complexity, and hence, a lower cost of implementation and subsequent operation. Such an approach is possible due to the use of error-correcting codes with low redundancy (significantly less than during duplication) dur-

ing organization of diagnostic systems [9]. Such codes, for example, include well-known classical and modified sum codes [10, 11].

A large class of codes is formed by sum codes during construction of which the weighting coefficients are attributed to separate bits of data vector. The principle of building such codes was proposed in the basic research [12] for the first time and studied in relation to the tasks of technical diagnostics in a small number of publications [13 – 16]. This scientific paper is devoted to the presentation of research results of the weighted sum codes regarding error detection at the outputs of the objects under diagnostics in the concurrent error-detection systems.

II. CLASSICAL AND WEIGHTED SUM CODES

Classical sum codes or Berger codes were proposed in 1961 and are built as follows [12]. The number of single bits (the weight r of the data vector is determined) is calculated in the data vector. The obtained number is represented in a binary form and recorded in the check vector bits. Therefore, the m -bit data vector requires $k = \lceil \log_2(m+1) \rceil$ check bits, where the record $\lceil \dots \rceil$ denotes an integer above the calculated value. Let's denote the classic sum code as the $S(m,k)$ -code, where m and k are the lengths of data vectors and check vectors.

The entire data vectors with the same weight have the same check vector in the $S(m,k)$ -code. The distribution of all data vectors between check vectors turns out to be uneven due to various amounts of vectors with different weights. This determines a high total number of errors undetectable via $S(m,k)$ -code. The formula for calculating a total number of errors undetectable via $S(m,k)$ -code is proposed in [13]:

$$N_{m,k} = \sum_{d=2}^l 2^{m-d} C_d^{\frac{d}{2}} C_m^d, \quad (1)$$

where d is the multiplicity of undetectable error (it can only be even for classical sum codes); the upper summation index $l = m$ with an even value of m and $l = m - 1$ with an odd value of m .

For example, when $m=6$, a total number of undetectable errors is to be determined by the value $N_{6,3} = 860$.

Errors that occur in the data vectors from the standpoint of their use in the discrete system technical diagnostics objectives are usually classified into four groups [18]. *Single* errors, i.e. the errors that occur during distortion of just one bit, are referred to the first group. Such errors should be detected via error-controlled codes. The second group is formed by *unidirectional* errors, i.e. the errors occurring during distortion of either zero or only single data bits. Single errors can also be considered as unidirectional single-bit errors. *Symmetric* errors form the third group of errors. Such errors occur during distortion of an equal quantity of zero and single data bits. On that basis, the symmetrical errors may have even multiplicity only. The fourth group is formed by *asymmetrical* errors that occur during distortion of the unequal number of zero and single data bits. Asymmetric errors may have multiples $d \geq 3$. The ratios between the total number of errors of various types are diverse with the different lengths of data vectors.

$S(m,k)$ -codes do not detect 100% of symmetric errors in the data vectors, whereas all other errors are detected by them. A vast number of methods for synthesis of testable discrete systems are based on this property of classical sum codes [19 – 26].

It was proved in [17] that the proportion of errors undetectable by $S(m,k)$ -codes of a specific even multiplicity of the total number of errors via a given multiplicity is a constant value being independent of the data vector length:

$$\beta_d = 100\% \cdot 2^{-d} C_d^{\frac{d}{2}}. \quad (2)$$

For example, $\beta_2 = 50\%$, $\beta_4 = 37.5\%$, $\beta_6 = 31.25\%$.

The formula (2) makes it possible to establish a very important disadvantage of $SM(m,k)$ -codes. These codes have a large quantity of undetectable errors with small multiplicities. This circumstance turns out to be significant while designing discrete systems, since it requires of the developer to make a significant redundancy into the structure of an object under diagnostics when solving technical diagnostics objectives.

The class of codes that is important for construction of discrete systems is formed by modular sum codes [18]. They are constructed as follows. The module M – some natural number is formed from the set $M \in \{2,3,\dots,m\}$. The weight r of the data vector is calculated. The smallest nonnegative deduction of the data vector weight – $r(\text{mod}M)$ number is determined based on a predetermined module. The resulting number is represented in a binary form and recorded in the check vector bits. We shall denote modular codes as $SM(m,k)$ -codes.

The number of check digits in $SM(m,k)$ -codes is determined by the selected module and may belong to the set $k \in \{1,2,\dots,\lceil \log_2(m+1) \rceil\}$. $SM(m,k)$ -codes are worse at detecting errors than classical sum codes. The total number of undetectable errors in $SM(m,k)$ -codes is determined from the formula [27]:

$$\begin{aligned} N_{m,k} &= \sum_{q=0}^{q=M-1} \left(\sum_{j=0}^{\lfloor \frac{m}{M} \rfloor} C_m^{jM+q} \right) \left(\sum_{j=0}^{\lfloor \frac{m}{M} \rfloor} C_m^{jM+q} - 1 \right) = \\ &= \sum_{q=0}^{q=M-1} A^2 \left(\sum_{j=0}^{\lfloor \frac{m}{M} \rfloor} C_m^{jM+q} \right), \quad jM + q \leq m. \end{aligned} \quad (3)$$

For example, we have 992 undetectable errors for $S4(6,2)$ -code. This number is 1.153 times more than the number of undetectable errors in the $S(6,3)$ code.

$SM(m,k)$ -codes have the following important properties [27]. Like the $S(m,k)$ -codes, they do not detect any symmetrical errors in the data vectors. Besides, unidirectional errors with multiplicities $d = jM$, $j = 1,2,\dots, \lfloor \frac{m}{M} \rfloor$ and asymmetrical errors with multiplicities $d = M + 2j$, $j = 1,2,\dots, \lfloor \frac{m-M}{2} \rfloor$ are not detected.

$SM(m,k)$ -codes, for which the module from the set $M \in \{2,4,\dots,2^{\lceil \log_2(m+1) \rceil - 1}\}$ was chosen during construction, have the property that they are most effective ones from the standpoint of error detection in comparison with all modular codes with the same value k . Besides, the functions describing their check bits coincide with the functions describing $\log_2 M$ of the lower order bits of the check vectors of the classical sum codes. This feature indicates that the complexity of the check equipment in discrete systems built based on modular sum codes will be less than when using classic sum codes for these purposes.

The general lack of classical and modular sum codes, consisting in a large number of undetectable errors, including those with small multiplicities, is eliminated by using the principle of weighting the data vector bits [12]. When constructing a weighted sum code or $WS(m,k)$ -code, the weighing coefficients in the form of natural numbers are assigned to the data vector bits, the weighing coefficients of the data vector unit bits are summed, and the resulting number is presented in a binary form and recorded into the check vector bits then.

In [16], when building a sum code, it is proposed to use a sequence of weighting coefficients from a series of increasing natural numbers with the exception of the powers of two. Such a $WS(m,k)$ code detects any single-bit errors or double-bit errors in data vectors and may be effectively used in data transferring and processing. However, such code turns out to be highly inefficient in the problems of discrete system synthesis, since it has a significant number of check bits, which affects the equipment duplication – it turns out to be comparable with duplication of data or components. In [28], it was proposed to choose the groups of weighting coefficients that form a series of increasing powers of 2 in order to build a code with detection of error bursts. Moreover, the size of the weighting coefficient group is determined by the multiplicity of the detected error burst. Such a weighted code has also a significant redundancy and is ineffective for the problems of the discrete system synthesis. In [13], it is proposed to choose an arbitrary sequence of weighting coefficients of data vector bits when building technical diagnostics systems, which makes it possible to influence on the resulting system structural redundancy and the features of failure detection within it. Such an approach is effective provided that the code redundancy will be insignificant compared to classical codes with the summation of single data bits.

As well as when building $SM(m,k)$ -codes, we can use modulo M residues when building weighted codes. For example, the application of such codes when solving problems of discrete system technical diagnostics is described in [14].

Thus, the sum codes may be classified as shown in Fig. 1.

An open issue remains with the choice of a sequence of bit weights coefficients in such a way that several conditions are to be provided. The first condition is to build simple code generator circuits. The second is to ensure the detection of a maximum number of faults with minimum equipment duplication. The third is with the receipt of such a code that will allow solving the problem of building fully self-checking structures. These are the issues that are considered in the following sections of the paper.

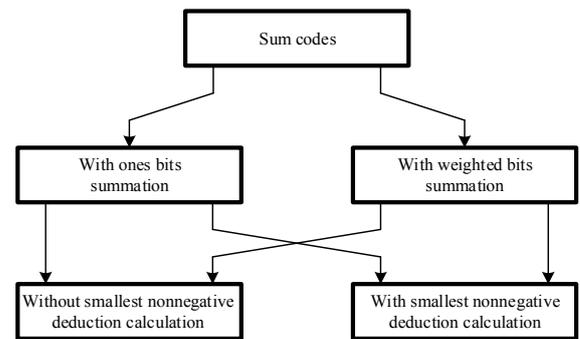


Fig. 1. Classification of sum codes.

III. CHARACTERISTICS OF ERROR DETECTION VIA WEIGHTED SUM CODES

Let's consider the features of error detection via $WS(m,k)$ -codes in case of their occurrence only in the data vector bits with the correctness of check vector bits. Such an assumption makes sense provided that the values of the data and check vectors bits are calculated via physically various devices [17]. The systems with built-on technical diagnostics equipment, for example, built-in check systems, or concurrent error-detection systems refer to such systems [4].

With an increase in the length of data vector from the value 2^q to the value $2^{q+1} - 1$ ($q \in \{2;3;\dots\}$) the number of various $WS(m,k)$ -codes decreases, since the number of methods for weighting data bits needed to satisfy the condition for maintaining the number of check bits equal to $k = \lceil \log_2(m+1) \rceil$ decreases:

$$m \leq W_{\max} \leq 2^{\lceil \log_2(m+1) \rceil} - 1, \quad (4)$$

where W_{\max} is the maximum value of data vector weight equal to the sum of all weighing coefficients w_i of data bits.

The tabular form of the task proposed in [17] is convenient in order to establish the characteristics of error detection via separable codes. In this form, the entire code data vectors are divided into check groups corresponding to all code check vectors. The content of errors undetectable via the code, their multiplicities and types are determined by analyzing the transitions possible in each control group.

The following approach may be used to determine the total number of errors undetectable via the code ($N_{m,k}$ value). When counting the number of undetectable errors in the $WS(m,k)$ -code, it is determined which quantity of data vectors makes it possible to obtain one or another total weight of the data vector. To do this, it is required to calculate the total number of various sums of data vector bit weighting coefficients.

Let's consider the counting procedure using the example of $WS(4,3)$ -code with a sequence of weighting coefficients $[1,1,2,3]$ (Tabl. 1).

The weight $W=0$ can only be obtained for the data vector $\langle 0000 \rangle$. The weight $W=1$ is obtained by using the vectors in which $f_4=1$ or $f_3=1$, and the remaining bits are equal to zero. The number of such cases is determined by the value C_2^4 (the number of combinations of two weighting coefficients one by one). The value $W=2$ can be obtained in several ways: either the sum of the weighting coefficients

$w_4 + w_3$ (C_2^1), or the use of the weighting coefficient w_2 (C_1^1). In other words, there is $C_2^1 + C_1^1$ of the option of obtaining $W=2$. Similarly, for $W=3$, we have the cases: $w_4 + w_2$, $w_3 + w_2$ and w_1 . The number of methods is the sum $C_2^1 + C_1^1$ (the number of different sums of unitary weighting coefficients and a coefficient equal to two, plus the only way of using a coefficient equal to three). For $W=4$, we have $C_3^1 + C_2^1$ (the sum $w_4 + w_3 + w_2$ and two options for summing the two coefficients – $w_4 + w_2$ and $w_3 + w_2$). For $W=5$, we obtain $C_3^1 + C_2^1$. For $W=6$, we obtain C_2^1 . For $W=7$, we obtain C_4^1 . The obtained variants of the weighting coefficient sums make it possible to count the total number of undetectable errors in the code by counting the sum of the number of undetectable errors in each check group (see Tabl. 1):

$$N_{m,k} = \sum_{j=1}^{W_{\max}-1} A_{p_j}^2, \quad (5)$$

where $A_{p_j}^2$ is the number of allocations from p_j by 2, and p_j is the number of options for formation of $W=j$ weight (the number of data vectors in the j -th control group).

TABLE I. $WS(4,3)$ -CODE WITH $[1,1,2,3]$

Weight W							
0	1	2	3	4	5	6	7
Data vectors							
0000	0100	0010	0001	0101	0011	0111	1111
	1000	1100	0110	1001	1101	1011	
			1010	1110			

For the example in question, we shall have:

$$\begin{aligned} N_{m,k} &= \sum_{j=1}^6 A_{p_j}^2 = A_{p_1}^2 + \dots + A_{p_6}^2 = \\ &= A_2^2 + A_2^2 + A_3^2 + A_3^2 + A_2^2 + A_2^2 = \\ &= 2 + 2 + 6 + 6 + 2 + 2 = 20. \end{aligned}$$

Analysis of the detection characteristics of the data vector errors via $WS(m,k)$ -codes shows that they refer to the codes with detection of any unidirectional errors in the data vectors.

Theorem 1. $WS(m,k)$ -codes detect any unidirectional errors in the data vectors.

Proof. Let's assume that the statement of Theorem 1 is false and there are unidirectional errors in the class undetectable via $WS(m,k)$ -codes. This means that there are data vectors at least in one control group of such a code, one of which contains at least two more unit bits more than the other one. For example, the vectors $\langle f_m f_{m-1} \dots f_i=1 f_{i+1}=1 \dots f_2 f_1 \rangle$ and $\langle f_m f_{m-1} \dots f_i=0 f_{i+1}=0 \dots f_2 f_1 \rangle$. The sums of the weighting coefficients of such vectors must coincide, inasmuch as they are located in the same check group. However, this is impossible, since the total weight of the first data vector is determined by the number

$$W_1 = f_m w_m + f_{m-1} w_{m-1} + \dots + 1 \cdot w_i + 1 \cdot w_{i+1} + \dots + f_2 w_2 + f_1 w_1,$$

and the total weight of the second data vector is determined via the number

$$W_2 = f_m w_m + f_{m-1} w_{m-1} + \dots + 0 \cdot w_i + 0 \cdot w_{i+1} + \dots + f_2 w_2 + f_1 w_1.$$

The first vector should have a weight greater than the second one by a value $w_i + w_{i+1}$. Thus, weighted sum codes will detect any unidirectional errors.

The theorem is proved.

A series of relative indicators is used for a comparative analysis of error detection characteristics via different sum codes. The proportion of errors undetectable via the code ($N_{m,k}$ value) of the total number of errors in the data vectors for a given length of the data vector (N_m value) is one of these indicators:

$$\gamma_m = 100\% \cdot \frac{N_{m,k}}{N_m}. \quad (6)$$

The lower the value γ_m , the more efficiently the code detects errors in the data vectors. The value $\gamma_m = 8.333\%$ for the $WS(4,3)$ -code with a sequence of weighting coefficients $[1,1,2,3]$ being considered. For comparison, $\gamma_m=22.5\%$ with a classical Berger code with the same length of data vector, which is 2.7 times more than the presented weighted code.

The so-called *efficiency coefficient* of using check bits via the code is another indicator for a sum code. This coefficient shows how close the sum code under consideration is to the code with the minimum total number of errors undetectable for m and k specific values ($N_{m,k}^{\min}$ of the errors is not detected via this code in data vectors):

$$\xi_{m,k} = 100\% \cdot \frac{N_{m,k}^{\min}}{N_{m,k}}. \quad (7)$$

The closer the coefficient $\xi_{m,k}$ to 100%, the closer the code under consideration is to the code with the minimum value of the total number of undetectable errors. For example, the value $\xi_{m,k}=80\%$ for $WS(4,3)$ -code with a sequence of weighting coefficients $[1,1,2,3]$.

The sum code for which $\xi_{m,k}=100\%$ is an *optimal code* upon criterion of the minimum of the total number of undetectable errors in the data vectors for specific values m and k . The optimal code should have a uniform distribution of the entire data vectors between all check vectors.

Theorem 2. There is no weighted sum code with a uniform distribution of the entire data vectors between the entire check vectors.

Proof. In order that $WS(m,k)$ -code to be optimal, it is necessary that per 2^{m-k} data vectors exactly to be contained in each of the 2^k check groups. Although, the principle of building the code makes it possible to obtain only one check vector with a weight $W=0$ and a weight $W_{\max} = w_1 + w_2 + \dots + w_{m-1} + w_m$. Check groups with these numbers contain only one data vector for any weighted code. As for the rest check groups of the $WS(m,k)$ -code, there are either no data vectors at all, or there are two or more data vectors. Such a code may not be optimal.

The theorem is proved.

The total number of undetectable errors in the weighted sum codes with the lengths of data vectors $m=4\div 15$ was calculated using the formula (5), which allowed us to determine the core indicators of error detection – γ_m and $\xi_{m,k}$. The change dependences in the values of the specified indicators for the $WS(4,3)$ and $WS(8,4)$ families are shown in Fig. 2 and Fig. 3. These codes are only particular examples, but they show common patterns inherent in the entire families of weighted codes, which manifest themselves during increas-

ing the lengths of data vectors with various combinations of weighted bits and weighing coefficient values.

The maximum number of weighted sum codes with the same number of check bits as that of the Berger code is built at the values $m = 2^p$, $m \in \{2;3;\dots\}$. For all other codes, with increase in the data vector length with a constant value of the check bit number, the number of weighting sequences is reduced allowing to build $WS(m,k)$ -codes with $k = \lceil \log_2(m+1) \rceil$.

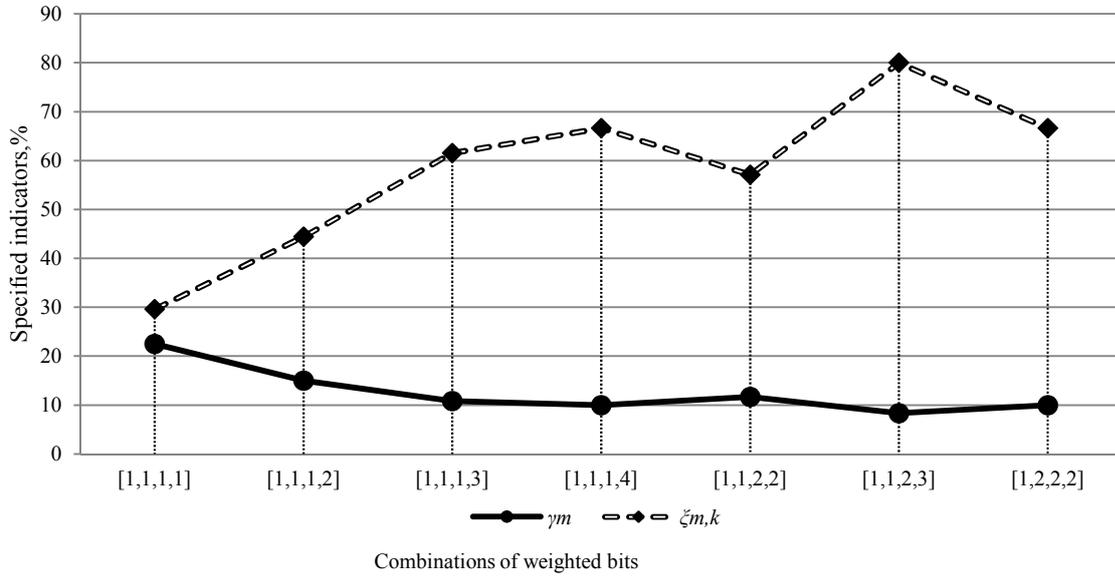


Fig. 2. Indicators of detection of the total number of errors via WS (4,3)-codes in the data vectors.

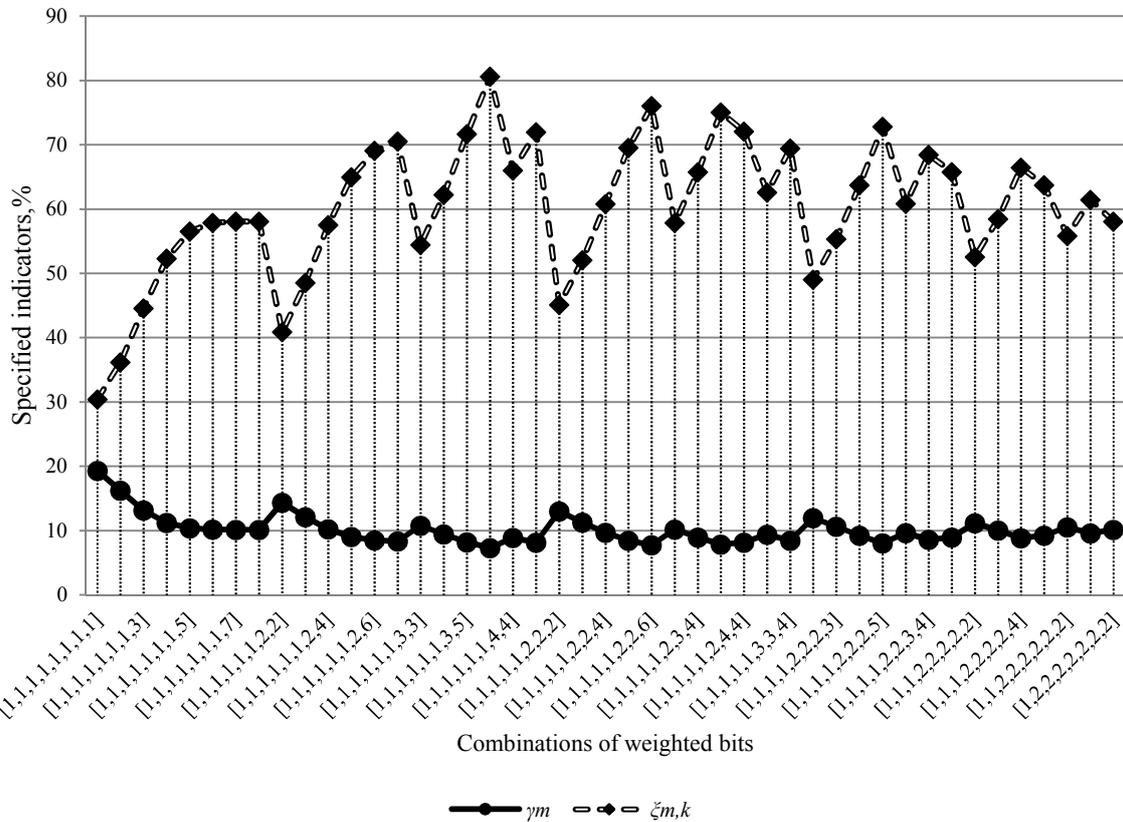


Fig. 3. Indicators of detection of the total number of errors via WS(8,4)-codes in the data vectors.

Let's analyze the data in Fig. 2 and Fig. 3. When weighing a single bit, the number of errors undetected by the code gradually decreases with increase of the weighting coefficient value reaching its maximum at $W_{\max} = 2^k - 1$. When weighing two bits by the same numbers, starting with $w_1=w_2=2$ and ending with the values $w_1=2^k - m - 1$ and $w_2 = 2$, a gradual decrease in the total number of errors undetected via the codes also takes place. Moreover, with the value W_{\max} , the code with two weighted digits at $w_1=2^k - m - 1$ and $w_2 = 2$ detects more errors than the code with one weighted digit at $w_1=2^k - m$. Then, when weighing two digits with the numbers $w_1=w_2=3$, with a further increase in the value of the weight w_1 , the characteristics are also improved. Generally, in the group of codes with the fixed values of weighting coefficients, as one of them increases, the improvements are observed in the detection of the total number of errors in the data vectors of the $WS(m,k)$ -codes.

IV. SYNTHESIS OF WEIGHTED SUM CODE GENERATORS

Since the sum operations are used during building a weighted code, it is most advisable to use a standard element base that includes full-adders (*FA*), half-adders (*HA*) and quarter-adders (*XOR*) for the implementation of code generators. The latter ones, however, will not be required during implementation of code generators for the construction of which no deductions are used based on a predetermined module. Structural flow charts of the simplest implementation variations of adders are shown in Fig. 4.

The adders fulfill the addition of signals entering on their inputs and form the values of obtained sums in a binary form at their outputs. A full adder operates with the data entering on three inputs, a half adder and a quarter-adder (adder in modulus two) – on two inputs. The output *S* of each device contains the sum of incoming numbers in the residue ring in modulus two, and the output *C* is designed to form a transfer signal.

The analysis of methods for building $WS(m,k)$ -code generators testified that the application of this approach [29] is the most effective:

1. The decomposition of weighting coefficients into sums of the number 2 powers is implemented.
2. The quantity of numbers of the *i*-th power of the number 2 – the N_i number is determined.
3. The value $i=0$ is set.
4. The *i*-th generator cascade, containing $\left\lfloor \frac{N_i - 1}{2} \right\rfloor$ full adders and $\frac{N_i - 1}{2} \pmod{2}$ half-adders is implemented.
5. The value *i* is increased by one: $i=i+1$.
6. The condition $i=i_{\max}?$ is checked. If yes, then the generator is built; if not, the next step is implemented.
7. The number of carry outputs of each adder of the *i*-1 cascade – $N_{C_{i-1}}$ number is determined.
8. The N_i number is corrected: $N_i = N_i + N_{C_{i-1}}$.
9. Steps 4 through 6 are repeated.

The structural flow chart of the $WS(8,5)$ -code generator with a sequence of $[w_8, w_7, w_6, w_5, w_4, w_3, w_2, w_1] = [1, 1, 2, 4, 3, 5, 2, 1]$ weighting coefficients is implemented in

Fig. 5 by way of example. The signals on all lines of the generator circuit are also shown when the data vector $\langle f_8, f_7, f_6, f_5, f_4, f_3, f_2, f_1 \rangle = \langle 10111101 \rangle$ arrives at the inputs.

To build a weighted code generator according to the above mentioned algorithm, the following decomposition of the weighting coefficients was fulfilled: $w_1 = w_7 = w_8 = 2^0$; $w_2 = w_6 = 2^1$; $w_3 = 2^2 + 2^0$; $w_4 = 2^1 + 2^0$; $w_5 = 2^2$. Thus, the quantity of different powers of the number 2 is equal to: $N_0 = 5$; $N_1 = 3$; $N_2 = 2$.

For the generator zero cascade performing the summation of bits with 2^0 weights, it took $\left\lfloor \frac{5-1}{2} \right\rfloor = 2$ full adders and $\left(\frac{5-1}{2} \right) \pmod{2} = 0$ half-adders. $\left\lfloor \frac{5}{2} \right\rfloor = 2$ adders totally. To build the first cascade of the generator, the number of carry outputs of each of the first cascade adders – $N_{C_0} = 2$ number should be added to $N_1 = 3$ number. Further, the bits with 2^1 weights are summed up, which requires $\left\lfloor \frac{(3+2)-1}{2} \right\rfloor = 2$ full adders and $\left(\frac{(3+2)-1}{2} \right) \pmod{2} = 0$ half adders. $\left\lfloor \frac{5}{2} \right\rfloor = 2$ adders totally. The second cascade of the generator is formed by one half adder and one full adder, since the numbers $N_2 = 2$ and $N_{C_1} = 2$ (the quantity of full adders equals $\left\lfloor \frac{(2+2)-1}{2} \right\rfloor = 1$, and the quantity of half adders equals $\left(\frac{(2+2)-1}{2} \right) \pmod{2} = 1$). The third cascade of the generator contains one half-adder fulfilling adding the carry output values of the second cascade adders.

It is worth noting that while decomposing weighting coefficients, the expansion in the maximum powers of number 2 is the most effective. For example, the number 5 should be decomposed as follows: $2^2 + 2^0$ (rather than so: $2^1 + 2^1 + 2^0$).

To assess the complexity of the structure obtained, the quantity of standard functional elements or the amount of inputs of internal logic elements may be used. The most convenient is the assessment of the complexity of the implementation of devices in terms of complexity of the library of standard functional elements *stdcell2_2.genlib* [30], which makes it possible to obtain a single relative number directly related to the sizes of the area occupied by the device on the chip. In the element library under consideration, the complexity of implementing a half adder is estimated by the value $L_{HA} = 72$ (two-input *XOR* element and two-input *AND* element), and the complexity of implementing a full adder is estimated by the value $L_{FA} = 176$ (two half-adders and two-input *OR* element). Taking this data into consideration, the complexity of technical implementation of $WS(8,5)$ -code generator with the sequence of weighting coefficients $[1, 1, 2, 4, 3, 5, 2, 1]$ is to be determined by the value $L_{WS} = 1024$.

The estimates of the technical implementation complexity of the generators of all $WS(8,5)$ -codes with the number of check bits $k = \lceil \log_2(m+1) \rceil$ are presented in Fig. 6, moreover, the codes are located in the sequence of increasing W_{\max}

number, which gives an idea of the regularities in changes of the implementation complexity indices. It should be noted that for 15 out of 44 weighted sum codes (except for the classical sum code), the implementation complexity is even less than the implementation complexity of the $S(8,4)$ -code generator according to the same method. Besides, as shown above, a much larger quantity of errors is detected via any weighted code rather than via $S(m,k)$ -code. As W_{\max} value increases, the number of undetectable errors decreases. At the same time, for each W_{\max} value, a sequence of weighting

coefficients may be chosen, which gives lesser implementation complexity compared to $S(m,k)$ -code.

Another element base may be used for implementation of $WS(m,k)$ code generators, for example, multiplexers [31]. The implementation of devices based on multiplexers is promising, since these devices are part of the logical blocks of modern programmable logical integrated circuits [32]. The simplest implementation variations of adders based on multiplexers with one and two address inputs are shown in Fig. 7.

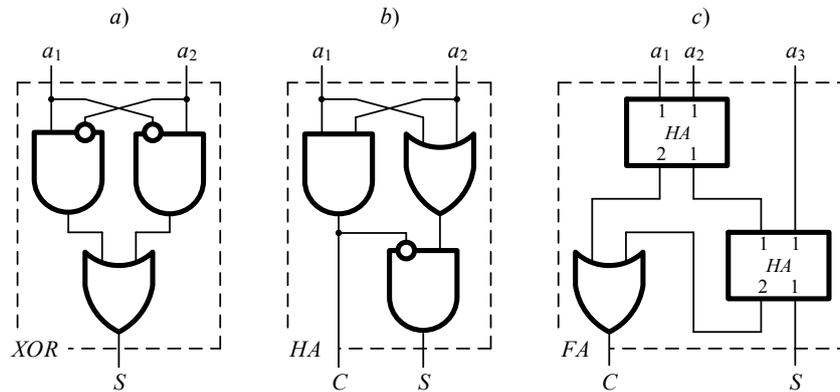


Fig. 4. Flow charts of adders.

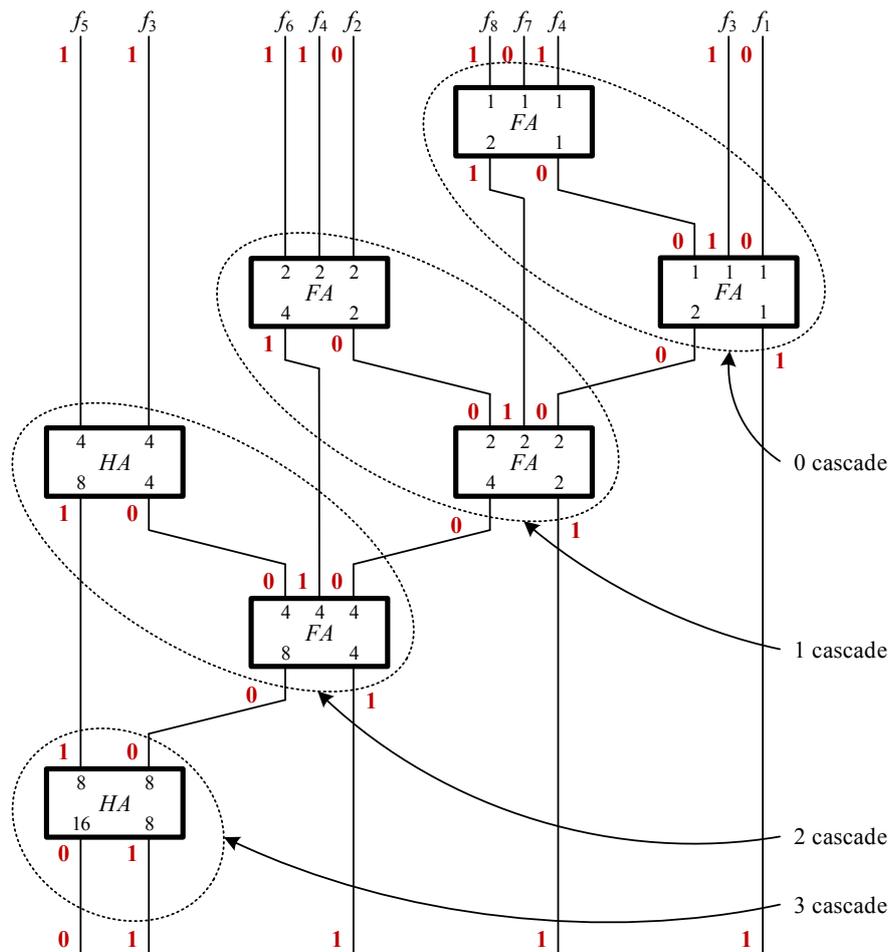


Fig. 5. $WS(8,5)$ -code generator with a sequence of weighting coefficients $[1,1,2,4,3,5,2,1]$ implemented based on standard adders.

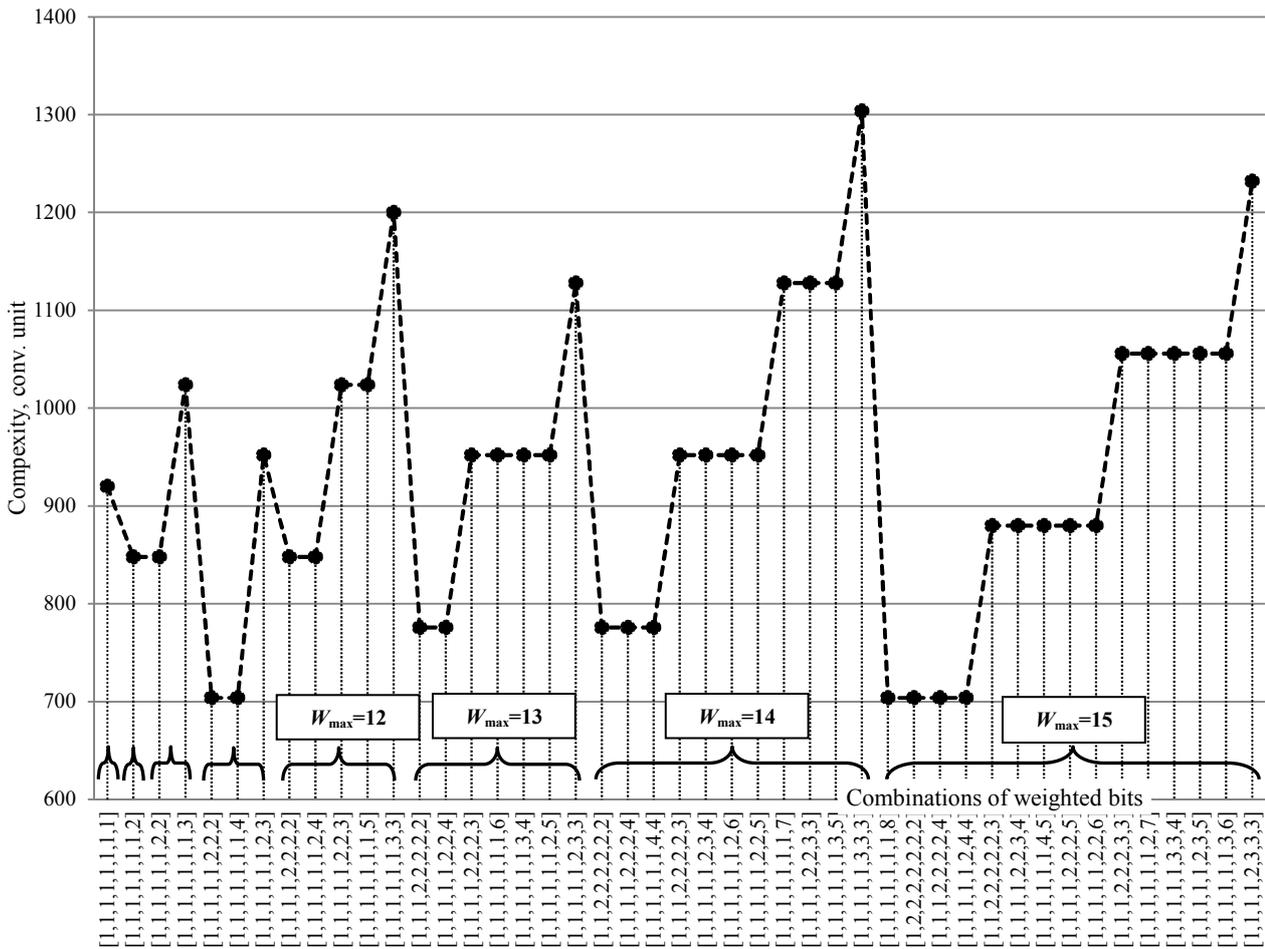


Fig. 6. Complexity of WS(8,4)-codes generators technical implementation.

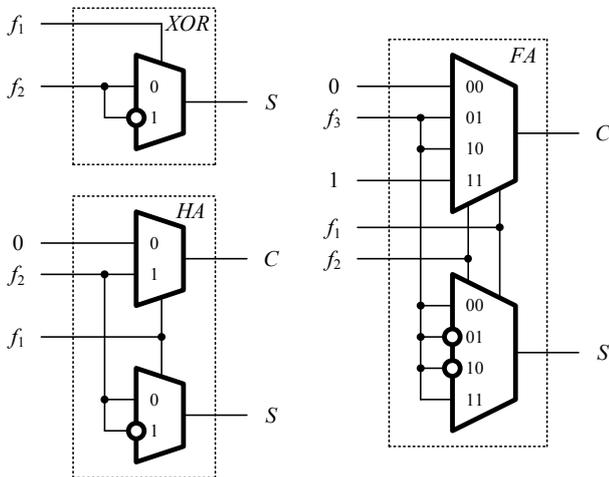


Fig. 7. Implementation of adders based on multiplexers.

To implement a weighted code generator on multiplexers, it is required to build it on the adders first using the algorithm proposed above, and then replace each adder with its standard implementation on multiplexers. The implementation of WS(8,5)-code generator with a sequence of weighting coefficients [1,1,2,4,3,5,2,1] on multiplexers is shown in Fig. 8.

V. CONCLUSION

Weighing the data vector bits is an effective procedure prior to building a sum code. It makes it possible to build sum codes that have a reduced total number of undetectable errors compared to classical Berger codes. At the same time, the weighted sum codes have a crucial property of Berger codes, which allows using them during construction of controllable discrete systems – unidirectional errors of any multiplicities are detected by them in data vectors. The presence of a certain proportion of asymmetrical errors in the class of undetectable ones is the price per preserving this property of a sum code with a reduced total number of undetectable errors. Weighted sum codes also detect symmetrical errors in data vectors much more efficiently than Berger codes. A similar conclusion may be drawn when comparing the detection characteristics via codes with the summation of double-bit errors in the data vectors. Such errors are the most frequent ones in discrete devices according to the statistics [33].

The synthesis method of weighted sum code generators proposed in the article allows us to build simple structures of these devices via modular connection of standard typical elements: adders or multiplexers. Besides, the proposed method makes us possible to implement weighted code generators based on a modern programmable element base with low hardware resource costs.

Due to development of the programmable element base and the ever-greater implementation of the control systems for critical technological processes implemented based on it in the entire fields of science and technology, the application of synthesis methods of controllable discrete devices is unquestionably important. The use of weighted sum codes in these tasks seems to be very effective.

REFERENCES

[1] R. Ubar, J. Raik, and H.-T. Vierhaus "Design and Test Technology for Dependable Systems-on-Chip (Premier Reference Source)",

Information Science Reference, Hershey – New York, IGI Global, 2011, 578 p.

[2] V. Kharchenko, Yu. Kondratenko, and J. Kacprzyk "Green IT Engineering: Concepts, Models, Complex Systems Architectures", Springer Book series "Studies in Systems, Decision and Control", Vol. 74, 2017, 305 p., doi: 10.1007/978-3-319-44162-7.

[3] V. Hahanov "Cyber Physical Computing for IoT-driven Services", New York, Springer International Publishing AG, 2018, 279 p., doi: 10.1007/978-3-319-54825-8.

[4] M. Nicolaidis "On-Line Testing for VLSI: State of the Art and Trends", Integration, 1998, Vol. 26, Issues 1-2, pp. 197-209, doi: 10.1016/S0167-9260(98)00028-5.

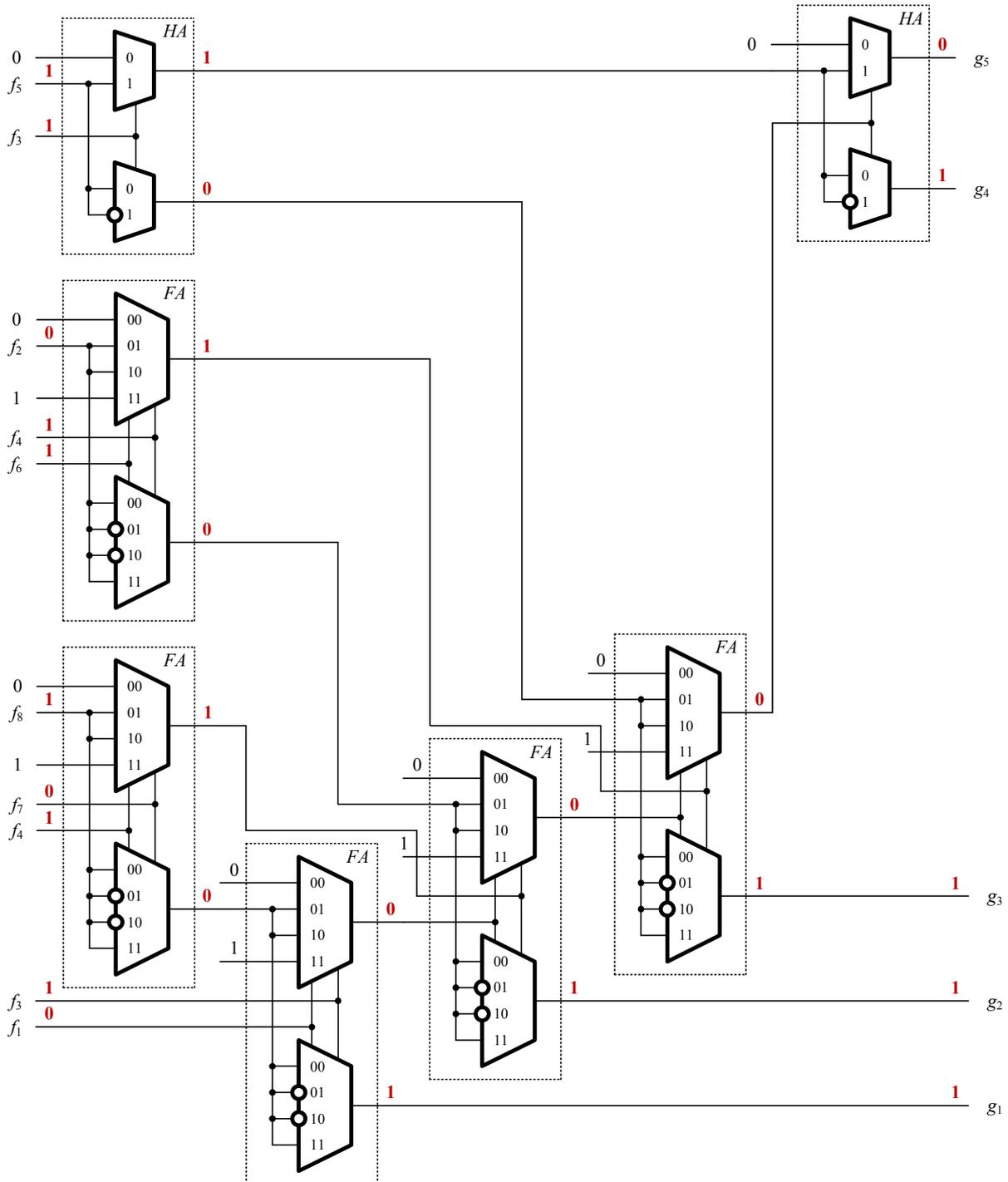


Fig. 8. WS (8,5)-code generator with a sequence of weight coefficients [1,1,2,4,3,5,2,1] implemented based on multiplexers.

- [5] R. Dobias, and H. Kubatova "FPGA Based Design of the Railway's Interlocking Equipments", Euromicro Symposium on Digital System Design (DSD 2004), 31 August – 3 September 2004, Rennes, France, pp. 467-473, doi: 10.1109/DSD.2004.1333312.
- [6] M. Gössel, V. Ocheretny, E. Sogomonyan, and D. Marienfeld "New Methods of Concurrent Checking: Edition 1", Dordrecht: Springer Science+Business Media B.V., 2008, 184 p.
- [7] P. Vít, J. Borecký, M. Kohlík, and H. Kubátová "Fault Tolerant Duplex System with High Availability for Practical Applications", 17th Euromicro Conference on Digital System Design, 27-29 August 2014, Verona, Italy, pp. 320-325, doi: 10.1109/DSD.2014.54.
- [8] G. Theeg, and S. Vlasenko "Railway Signalling & Interlocking: 2nd Edition", Germany, Hamburg: PMC Media House GmbH, 2018, 458 p.
- [9] E. Fujiwara "Code Design for Dependable Systems: Theory and Practical Applications", John Wiley & Sons, 2006, 720 p.
- [10] S.J. Piestrak "Design of Self-Testing Checkers for Unidirectional Error Detecting Codes", Wrocław: Ofiyna Wydawnicza Politechniki Wrocławskiej, 1995, 111 p.
- [11] D. Efanov, V. Sapozhnikov, and VI. Sapozhnikov "Generalized Algorithm of Building Summation Codes for the Tasks of Technical Diagnostics of Discrete Systems", Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 365-371, doi: 10.1109/EWDTS.2017.8110126.
- [12] J.M. Berger "A Note on Error Detection Codes for Asymmetric Channels", Information and Control, 1961, Vol. 4, Issue 1, pp. 68-73, doi: 10.1016/S0019-9958(61)80037-5.
- [13] D. Das, and N.A. Touba "Weight-Based Codes and Their Application to Concurrent Error Detection of Multilevel Circuits", Proceedings of the 17th IEEE VLSI Test Symposium, USA, CA, Dana Point, April 25-29, 1999, pp. 370-376.
- [14] D. Das, N.A. Touba, M. Seuring, and M. Gossel "Low Cost Concurrent Error Detection Based on Modulo Weight-Based Codes", Proceedings of IEEE 6th International On-Line Testing Workshop (IOLTW), Spain, Palma de Mallorca, July 3-5, 2000, pp. 171-176, doi: 10.1109/OLT.2000.856633.
- [15] V. Sapozhnikov, VI. Sapozhnikov, D. Efanov, and D. Nikitin "Combinational Circuits Checking on the Base of Sum Codes with One Weighted Data Bit", Proceedings of 12th IEEE East-West Design & Test Symposium (EWDTS'2014), Kyev, Ukraine, September 26-29, 2014, pp. 126-136, doi: 10.1109/EWDTS.2014.7027064.
- [16] D. Efanov, V. Sapozhnikov, VI. Sapozhnikov, and D. Nikitin "Sum Code Formation with Minimum Total Number of Undetectable Errors in Data Vectors", Proceedings of 13th IEEE East-West Design & Test Symposium (EWDTS'2015), Batumi, Georgia, September 26-29, 2015, pp. 141-148, doi: 10.1109/EWDTS.2015.7493112.
- [17] D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "On Summation Code Properties in Functional Control Circuits", Automation and Remote Control, 2010, Vol. 71, Issue 6, pp. 1117-1123, doi: 10.1134/S0005117910060123.
- [18] V. Sapozhnikov, VI. Sapozhnikov, and D. Efanov "Modular Sum Code in Building Testable Discrete Systems", Proceedings of 13th IEEE East-West Design & Test Symposium (EWDTS'2015), Batumi, Georgia, September 26-29, 2015, pp. 181-187, doi: 10.1109/EWDTS.2015.7493133.
- [19] E.S. Sogomonyan, and M. Gössel "Design of Self-Testing and On-Line Fault Detection Combinational Circuits with Weakly Independent Outputs", Journal of Electronic Testing: Theory and Applications, 1993, Vol. 4, Issue 4, pp. 267-281, doi: 10.1007/BF00971975.
- [20] F.Y. Busaba, and P.K. Lala "Self-Checking Combinational Circuit Design for Single and Unidirectional Multibit Errors", Journal of Electronic Testing: Theory and Applications, 1994, Vol. 5, Issue 1, pp. 19-28, DOI: 10.1007/BF00971960.
- [21] V.V. Saposhnikov, A. Morosov, VI.V. Saposhnikov, and M. Gössel "A New Design Method for Self-Checking Unidirectional Combinational Circuits", Journal of Electronic Testing: Theory and Applications, 1998, Vol. 12, Issue 1-2, pp. 41-53, doi: 10.1023/A:1008257118423.
- [22] A. Morosow, V.V. Sapozhnikov, VI.V. Sapozhnikov, and M. Goessel "Self-Checking Combinational Circuits with Unidirectionally Independent Outputs", VLSI Design, 1998, Vol. 5, Issue 4, pp. 333-345, doi: 10.1155/1998/20389.
- [23] A.Yu. Matrosova, I. Levin, and S.A. Ostanin "Self-Checking Synchronous FSM Network Design with Low Overhead", VLSI Design, 2000, Vol. 11, Issue 1, pp. 47-58, doi: 10.1155/2000/46578.
- [24] A. Matrosova, and E. Mitrofanov "Pseudo-Exhaustive Testing of Sequential Circuits for Multiple Stuck-at Faults", Proceedings of 14th IEEE East-West Design & Test Symposium (EWDTS'2016), Yerevan, Armenia, October 14-17, 2016, pp. 533-536, doi: 10.1109/EWDTS.2016.7807694.
- [25] D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "Conditions for Detecting a Logical Element Fault in a Combination Device under Concurrent Checking Based on Berger's Code", Automation and Remote Control, 2017, Vol. 78, Issue 5, pp. 891-901, doi: 10.1134/S0005117917040113.
- [26] S. Ostanin "Self-Checking Synchronous FSM Network Design for Path Delay Faults", Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 696-699, doi: 10.1109/EWDTS.2017.8110129.
- [27] D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "Application of Modular Summation Codes to Concurrent Error Detection Systems for Combinational Boolean Circuits", Automation and Remote Control, 2015, Vol. 76, Issue 10, pp. 1834-1848, doi: 10.1134/S0005117915100112.
- [28] J.M. Berger "A Note on Burst Detection Sum Codes", Information and Control, 1961, Vol. 4, Issue 2-3, pp. 297-299, doi: 10.1016/S0019-9958(61)80024-7.
- [29] M. Kang "A Study of Self-Checking Circuit Design Based on Berger Code" (in Chinese), Master's dissertation, Harbin Engineering University, 2007, 64 p.
- [30] E.M. Sentovich, K.J. Singh, C. Moon, H. Savoj, R.K. Brayton, and A. Sangiovanni-Vincentelli "Sequential Circuit Design Using Synthesis and Optimization", Proceedings IEEE International Conference on Computer Design: VLSI in Computers & Processors, 11-14 October 1992, Cambridge, MA, USA, USA pp. 328-333, doi: 10.1109/ICCD.1992.276282.
- [31] C. Maxfield "The Design Warrior's Guide to FPGA's: Devices, Tools and Flows", Boston: Newnes, 2004, 542 p.
- [32] D.M. Harris, and S.L. Harris "Digital Design and Computer Architecture", Morgan Kaufmann, 2012, 569 p.
- [33] V. Sapozhnikov, D. Efanov, VI. Sapozhnikov, and V. Dmitriev "Method of Combinational Circuits Testing by Dividing its Outputs into Groups and Using Codes, that Effectively Detect Double Errors", Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 129-136, doi: 10.1109/EWDTS.2017.8110123.

Algorithm for Extraction of the Iris Region in an Eye Image

Sh.Kh. Fazilov

Scientific and Innovation Center, Information and Communication Technologies
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Tashkent, Uzbekistan
sh.fazilov@mail.ru

O.R. Yusupov

Department of Mathematical Modelling
Samarkand State University
Samarkand, Uzbekistan
ozodyusupov@gmail.com

Abstract—The iris recognition system depends heavily on the accuracy of the selected region of the iris. The paper considers the task of selecting the iris region in an eye image. To solve the problem, an algorithm has been developed that allows selecting the iris inner and outer boundaries in an acceptable time. The algorithm differs from the known ones by resistance to clutters in the form of glare, eyelashes, shadows and eyelids, as well as by high-speed operation, and allows working with images recorded under different conditions. Static results are given that determine the accuracy of the results and the time of the operation algorithm. To analyze the effectiveness of the proposed algorithms, the CASIA V4 database consisting of 52034 images of various positions and quality was used.

Keywords—image processing, biometric technology, binarization, iris, Hough transform, Sobel operator

I. INTRODUCTION

Current trends in the field of data protection dictate strict requirements imposed to authorization systems for users of information resources. This gave rise to the widespread use of biometric identification, which allows a substantial increase in the level of information security.

One of promising biometric technologies that has been developed recently is iris recognition. The shape of the iris segments is a stable, well-pronounced and informative biometric feature. The iris is located on the front of the eyeball and has an almost annular form. It is about 11 millimeters in size. The form and size of the iris outer boundary are constant (they do not change over time) and almost the same for all people. The iris inner boundary is given by the pupil located approximately in its center. In the first approximation, the iris inner and outer boundaries can be considered concentric circles (Fig. 1).

Building recognition systems using an iris image is carried out according to the classical scheme: the extraction of an informative region (the iris) in the image, the formation of an attribute description of the extracted region, the comparison of the formed attribute descriptions. The above scheme can include the steps of assessing the quality of both the obtained images and the work of the individual stages of the method [1]. It should be noted that an important role in achieving high quality indicators of iris recognition is played by the accuracy of the first step - the extraction of the iris inner and outer boundaries in the image.

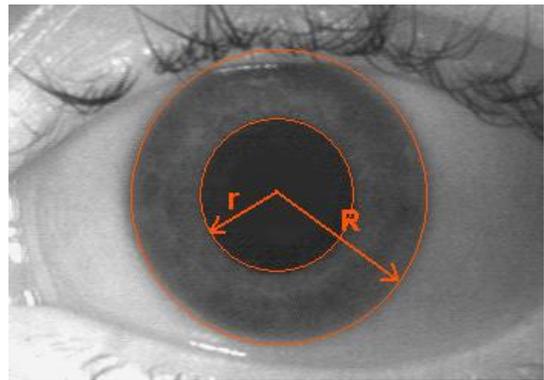


Fig. 1. The iris boundaries

The extraction of the iris inner and outer boundaries may be incorrectly performed for the following reasons: different lighting conditions when registering with different systems; eyelashes and eyelids covering the iris; the pupil defects; the iris dark colour (the pupil and the iris pattern are difficult to distinguish); a chronically dilated pupil (the pupil diameter should be less than 75% of the iris diameter); glare on the iris; head movement, blinking, inaccurate positioning of the head; changes in the iris caused by changes in the pupil size or shape, etc.

Many algorithms have been developed and are being applied to solve the problem of extracting an area in the image that corresponds to the iris. The classical method proposed in [2] uses an integro-differential operator, which selects radially symmetric structures and has high accuracy and stability, but its computational complexity is unacceptable for most applications [3]. Attempts have been made to reduce its computational complexity, for example, through pyramidal processing [4].

Various radial filters in the framework of solving this problem were considered in [5, 6] and others. A large number of methods are based on binarization of the image with the subsequent extraction of a single [7] or the most appropriate [8] object as a pupil. These methods show good results in images with dark pupils [9], but they are ineffective on other types of images [10]. The morphological methods of the pupil detection show the same performance. Both simple methods, for example, determining the center of a pupil as a point furthest from the bright regions [11] and sophisticated ones [12] essentially rely

on the fact that the pupil is the darkest or at least one of the darkest objects in the image. An interesting extension here is the possibility of special lighting and the use of the red-eye effect [13].

Another significant class of methods is the use of various types of the circle Hough transform: from direct construction of a three-dimensional (two center coordinates and a radius) accumulator, as suggested in the classic work [14], to complex methods using gradients [15], the pre-processing of an image with the selection of its regions by clustering methods [16], transformations with separated accumulators [17]. In such methods, as the first step, a gradient image transformation, i.e. an approximate calculation of the partial derivatives of the image brightness at each point, is performed. Since the pupil is very different from the surrounding iris in brightness, the gradient values are large at its boundary - the circumference. Then, it is necessary to find the parameters of this circumference, for which the Hough transform [18], which allows finding the parameters of curves of a given type (in this case, the circumferences), is used. These are the methods used in [19] and [20]. In [21], the pre-binarization of the image is also used for acceleration. In [22], the equivalence to the circle Hough transform is shown for the application of a certain operator. Various methods have been developed to reduce the computational complexity of the problem. Many other approaches have been developed: obtaining the pupil circumference as an escribed circle for sets of three points (triangulation) [23], using active contours [24], classifiers, including Adaboost [25], the support vector machine [26], multiple-scale processing [27], including wavelets [28], combinations of several methods (for example, based on the extraction of regions and the extraction of boundaries) [29]. In [8], an individual method is used to find the coordinates of the center of an eye (without determining the pupil radius) only, and it is performed approximately to be further refined by other methods.

II. STATEMENT OF THE PROBLEM

The purpose of this paper is to develop an algorithm that would provide the extraction of inner and outer boundaries in an iris image, with indicators of accuracy, reliability and speed of work that would be acceptable for practical use in person identification systems, be highly resistant to various interference and work in low-quality images.

The input data of the algorithm is a monochrome bitmap of the iris with the size of $M \times N$ pixels.

The output data of the algorithm is the parameters of two circles approximating the boundaries of the inner and outer boundaries, namely the coordinates of the centers and radii (ζ_p, ξ_p, r_p) and (ζ_i, ξ_i, r_i) .

III. DESCRIPTION OF THE ALGORITHM

Two stages are required to solve the problem. The iris inner boundary shall be found at the first stage, and its outer boundary shall be found at the other one.

A. The algorithm for extraction of the iris inner boundary

To solve this problem, an algorithm has been developed that allows localizing the pupil in the image of an eye in a reasonable time. The algorithm consists of the following steps.

Step 1. The results of selecting the iris inner boundary may be incorrect due to noise in the image, so it is important to filter out noise to prevent false detections caused by noise. To reduce the noise contained in the original image, as well as to remove high-frequency components from the original image in order to carefully examine the content of low-frequency components, the two-dimensional Gaussian filter is used in the form

$$G(\zeta, \xi) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{\zeta^2 + \xi^2}{2\sigma^2}}.$$

The Gaussian filter leaves the low-frequency components of the image intact and attenuates the high-frequency components, which results in blurred images. When smoothing the image using the Gaussian filter, it is possible to use masks with different weights. The larger σ , the more the image is blurred when applying the Gaussian filter. Therefore, the value of these parameters is selected depending on the degree to which an image needs to be blurred.

Step 2. The image shall be divided into rectangles, in each of which the average brightness is determined by the formula

$$P_{k,l} = \frac{1}{v^2} \sum_{\alpha=1}^v \sum_{\beta=1}^v I(k \cdot v + \alpha, l \cdot v + \beta),$$

where

$$v = \min\left(\frac{M}{r_{pmin}}, \frac{N}{r_{pmin}}\right), \quad M_1 = \left\lceil \frac{M}{v} \right\rceil, \quad N_1 = \left\lceil \frac{N}{v} \right\rceil,$$

$$k = 0..M_1 - 1, \quad l = 0..N_1 - 1$$

Further, it is considered that the rectangle with the minimum brightness lies in the region of the pupil and is located $P_{min} = \min_{k,l} (P_{k,l})$.

Step 3. The image is binarized with a certain threshold, after which the components, on which the pupil is present, remain. Binarization is the division of all pixels of an image into two classes according to a certain brightness threshold τ . The value of zero shall be assigned to pixels with the brightness that is lower than τ , and the value of one shall be assigned to pixels with the brightness that is higher than τ :

$$I_b(\zeta, \xi) = \begin{cases} 1, & I(\zeta, \xi) \leq \tau; \\ 0, & I(\zeta, \xi) > \tau. \end{cases}$$

The threshold $\tau = \eta \cdot P_{min}$ where $\eta = 1.3$ is the coefficient that depends on the image brightness, $P_{min} = \min_{k,l} (P_{k,l})$ is the

average brightness in the minimum brightness rectangle defined at the previous step.

Step 4. To select the pupil boundary, the Sobel operator is used. The Sobel operator calculates the brightness gradient of an image in each pixel. On the other hand, the approximation of the gradient used by it is rather rough, this especially affects the high-frequency oscillations of the image. At first, the gradient is estimated along the directions of the vertical and horizontal axis. This is performed using two kernels:

$$U_{G_\zeta} = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \quad U_{G_\xi} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix},$$

$$G_\zeta = U_{G_\zeta} * I_b, \quad G_\xi = U_{G_\xi} * I_b,$$

where * is the two-dimensional convolution operation

The magnitude of the gradient is usually defined as follows:

$$I_{Grad} = \sqrt{G_\zeta^2 + G_\xi^2}.$$

After determining G_{max} , to extract the desired contour, the following shall be found

$$G_{max} = \max_{\zeta, \xi} (I_{Grad}(\zeta, \xi)).$$

The extraction of the desired contour was carried out as follows:

$$I_c(\zeta, \xi) = \begin{cases} 1, & I_{Grad}(\zeta, \xi) \leq \gamma; \\ 0, & I_{Grad}(\zeta, \xi) > \gamma, \end{cases}$$

where $\gamma = \eta_1 \cdot G_{max}$, with η_1 being determined experimentally (our case $\eta_1 = 0.5$).

Step 5. To find the circle circumscribing the pupil, the following kind of Hough transform is used

$$H(\zeta_0, \xi_0, r) = \sum h(\zeta_i, \xi_i, \zeta_0, \xi_0, r),$$

$$h(\zeta_i, \xi_i, \zeta_0, \xi_0, r) = \begin{cases} 1, & (\zeta_i - \zeta_0)^2 + (\xi_i - \xi_0)^2 = r^2; \\ 0, & \text{otherwise,} \end{cases}$$

where (ζ_0, ξ_0) are the coordinates of the pupil center to be determined; (ζ_i, ξ_i) is a pixel in the image from a certain neighborhood; $r \in [r_{pmin}; r_{pmax}]$ are possible radii of the circle.

The developed algorithm significantly simplifies the search task and gives good results of the accuracy of determining the inner boundary, allows reducing the dimension of the parameter space in comparison with methods that determine both the center

and radius, increases the stability of the method, especially in images with a noisy pupil image.

B. The algorithm for the extraction of the iris outer boundary

Step 1. The approximate radius r_i of the iris is selected from $r_i \in [r_{imin}; r_{imax}]$ so are the coordinates of the center of its inner boundary (ζ_p, ξ_p) .

Step 2. From the center of inner boundary, moving left, right, up and down at a distance ℓ_{dis} ($\ell_{dis} = 1.2 \cdot r_i$), a rectangular area is selected from the eye original image.

Step 3. A Gaussian filter is applied to smooth the selected image:

$$G(\zeta, \xi) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{\zeta^2 + \xi^2}{2\sigma^2}}.$$

Step 4. To remove noise from the selected image, a median filter is applied in the form:

$$I_m(\zeta, \xi) = \text{med}_{(s,t) \in S_{\zeta\xi}} \{I_g(s,t)\},$$

where $S_{\zeta\xi}$ is the rectangular area of the image I_g with the size of $k \times k$ pixels.

The median filter copes well with low to moderate noise levels, but to suppress more intense noise, it is necessary to use a median filter with a larger window of filtration. However, an increase in the window size results in the growth of the median component ability and can lead to distortion of the object outlines. In addition, small objects can be removed entirely from the image. Therefore, in each case, the filter parameters are adjusted depending on the degree of distortion and the characteristic size of the observed objects.

Step 5. To improve the image quality, gamma correction is applied in the form

$$I_\gamma(\zeta, \xi) = c \cdot I_m^\gamma(\zeta, \xi)$$

where c and γ are constants.

Step 6. The iris edge points I_γ are found using the Canny edge detector [30].

Step 7. The pupil image is cleared from the edge points and the eyelid outer and inner boundaries.

Step 8. To find the circle circumscribing the pupil, the following type of the Hough transform is used

$$H(\zeta_0, \xi_0, r) = \sum h(\zeta_i, \xi_i, \zeta_0, \xi_0, r),$$

$$h(\zeta_i, \xi_i, \zeta_0, \xi_0, r) = \begin{cases} 1, & (\zeta_i - \zeta_0)^2 + (\xi_i - \xi_0)^2 = r^2; \\ 0, & \text{otherwise,} \end{cases}$$

where (ζ_0, ξ_0) are the coordinates of the outer boundary to be determined; (ζ_r, ξ_r) is a pixel in the image from a certain neighborhood; $r \in [r_{\min}, r_{\max}]$ are possible radii of the circle.

The developed extraction algorithm has the following advantages over the known ones: as the basis for the operation of the algorithm, the results of the selected iris inner boundary, rather than the original image, are used, which greatly simplifies the process of extracting the outer boundary and increases the high-speed operation of the algorithm; ensures the simultaneous elimination of noise and iris defects and enhances image contrast, which reduces the quantity of the boundary extraction stages.

Fig. 2 shows the results of extracting the iris region using the proposed algorithms. It can be seen in the figure that the iris inner and outer boundaries are sufficiently selected regardless of the presence of eyelids, eyelashes, glares, and other interferences.

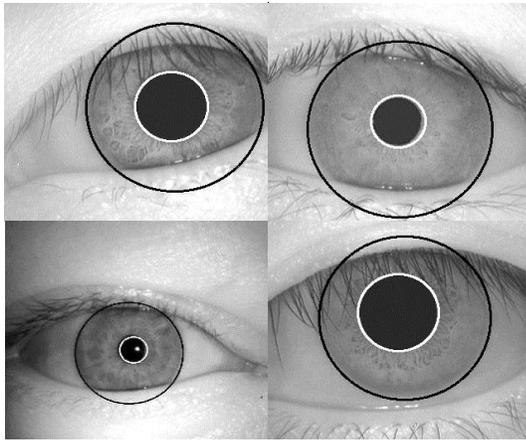


Fig. 2. Examples of the iris region extraction

IV. EXPERIMENTS

The purpose of a computational experiment is to verify the efficiency of the proposed algorithm on real databases and to compare the accuracy of the results and the time taken by the algorithm with those of well-known Daugman's [2] and Wildas [1] algorithms.

The experiment was conducted on a computer with the following characteristics: the processor type was Intel Core i5-2430M, the processor speed was 2.4 GHz, the operating system type was Windows 8.1, the RAM was 6 Gb. All algorithms are implemented in the Matlab R2017a system.

The CASIA-Iris V4 [31] database was chosen as the basis for testing the proposed algorithms. All images were recorded in the near infrared range, with the colour depth of 8 bits, JPEG file format. The CASIA-Iris V4 database is divided into six groups of images: CASIA-Iris-Interval, CASIA-Iris-Twins, CASIA-Iris-Thousand, CASIA-Iris-Lamp, CASIA-Iris-Syn, and CASIA-Iris-Distance. Only the first five groups of images were used in the experiment.

CASIA-Iris-Interval was obtained in two sessions and contains 249 objects, 395 classes, 2639 images. The number of images per eye is 1 to 26. Images of this database were shot with

our own device in the near infrared range with a resolution of 320×280 pixels. In the group of images, there are ones with strong occlusion and poor contrast. The image quality is poor, the iris texture is fuzzy, there are clutters on the pupil in the form of horizontal stripes; the iris boundary and the pupil are difficult to distinguish.

CASIA-Iris-Lamp was obtained in one session and contains 411 objects, 819 classes, 16212 images. The number of images per eye is 10 to 20. The size of an image is 640×480 pixels. The images were recorded in the room with an OKI(IRISPASS-h) manual device. The image quality is visually better. A large portion of the images has a significant shading of the iris with eyelids and eyelashes.

CASIA-Iris-Twins is the first database obtained in one session and contains images of the eyes of twins. The size of an image is 640×480 pixels; the database contains 200 objects, 400 classes, 3183 images. The number of images per eye is 5 to 10. The images were recorded under different lighting conditions. The visibility of the iris is poor.

CASIA-Iris-Thousand is the first database that contains more than a thousand objects obtained in one session. It contains 100 objects, 200 classes, 20000 images. The number of images per eye is 10. The size of an image is 640×480 pixels.

CASIA-Iris-Syn is the database containing synthesized images of the iris obtained in one session. It contains 1000 objects, 1000 classes, 10000 images. The size of an image is 640×480 pixels.

TABLE I. THE RESULTS OF THE WORK OF THE ALGORITHMS FOR THE IRIS INNER BOUNDARY

Algorithm	Accuracy	Average operating time, sec
Daugman's	96.2%	0.495
Wildas	95.7%	0.553
Proposed	98.1%	0.283

TABLE II. THE RESULTS OF THE WORK OF THE ALGORITHMS FOR THE IRIS OUTER BOUNDARY

Algorithm	Accuracy	Average operating time, sec
Daugman's	95.6%	0.596
Wildas	94.8%	0.693
Proposed	97.2%	0.387

During the experiment, 52034 images were used. As can be seen from Tables 1 and 2, the proposed algorithms spend on the average one image window of 0.283 s to determine the iris inner boundary and of 0.387 s to determine the iris outer boundary, which is almost 2 times less than the same indicator of the compared algorithms.

V. CONCLUSION

An algorithm of searching for the iris inner and outer boundaries in the image that is efficient in terms of speed and accuracy is proposed. A computational experiment was conducted to test the working efficiency of the proposed algorithm in real images obtained from the CASIA data set. A comparative analysis of the proposed algorithm with the Daugman's and Wildas algorithms in terms of accuracy and operating time was conducted. The results of the comparison

showed that the proposed algorithm has advantages in terms of accuracy as compared to known algorithms.

REFERENCES

- [1] R. P. Wildes, "Iris recognition: an emerging biometric technology," *Proceedings of the IEEE*, vol. 85, pp. 1348–1363, 1997.
- [2] J. Daugman, "How iris recognition work," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 21–30, 2004.
- [3] N. Barzegar, and M. S. Moin, "A New Approach for Iris Localization in Iris Recognition Systems," *Proceedings IEEE/ACS International Conference on Computer Systems and Applications*, pp. 516–523, 31 March - 4 April 2008.
- [4] M. Shamsi, P. Saad, S. Ibrahim, and A. R. Kenari, "Fast Algorithm for Iris Localization Using Daugman Circular Integro-Differential Operator," *International Conference of Soft Computing and Pattern Recognition*, pp. 393–398, 2009.
- [5] F. Alonso-Fernandez, and J. Bigun, "Iris Segmentation Using the Generalized Structure Tensor," *SSBA Symposium*, 8-9 March, 2012.
- [6] D. S. Jeong, J. W. Hwang, B. J. Kang, K. R. Park, C. S. Wonc, D. K. Park, and J. H. Kim, "A new iris segmentation method for non-ideal iris images," *Image and Vision Computing*, vol. 28, pp. 254–260, 2010.
- [7] T. Maenpaa, "An Iterative Algorithm for Fast Iris Detection," *Proceedings IWBRIS*, pp. 127-134, 2005.
- [8] L. Pan, M. Xie, and Z. Ma, "Iris Localization based on Multi-resolution Analysis," *19th International Conference on Pattern Recognition*, pp. 1-4, 2008.
- [9] D. Chen, J. Bai, and Z. Qu, "Research on Pupil Center Location Based on Improved Hough Transform and Edge Gradient Algorithm," *National Conference on Information Technology and Computer Science*, pp. 47-51, 2012.
- [10] P. Phillips, W. Scruggs, A. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "Frvt2006 and ice2006 large-scale experimental results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 32, pp. 831-846, 2010.
- [11] B. Lipinski, "Iris Recognition: Detecting the pupil," 2004, available at <http://cnx.org/content/m12487/latest/>.
- [12] F. Gui, and L. Qiwei, "Iris localization scheme based on morphology and gaussian filtering," *IEEE Conference on Signal-Image Technologies and Internet-Based System*, pp. 798-803, 2007.
- [13] C. H. Morimoto, T. T. Santos, and A. S. Muniz, "Automatic Iris Segmentation Using Active Near Infra Red Lighting," *18th Brazilian Symposium on Computer Graphics and Image Processing*, pp. 37-43, 2005.
- [14] R. P. Wildes, J. C. Asmuth, K. J. Hanna, S. C. Hsu, R. J. Kolczynski, J. R. Matey, and S. E. McBride, "Automated, non-invasive iris recognition system and method," *US Patent no. 5751836*, 1998.
- [15] A. Basit, and M. Y. Javed, "Localization of iris in gray scale image using intensity gradient," *Opt. Lasers Eng.*, vol. 45, pp. 1107–1114, 2007.
- [16] H. Proenca, "Iris Recognition: A Method To Segment Visible Wavelength Iris Images Acquired On-The-Move and At-A-Distance," *ISVC 2008, Part I, LNCS vol. 5358*, pp. 731–742, 2008.
- [17] D. E. Benn, M. S. Nixon, and J. N. Carter, "Robust eye centre extraction using the Hough transform," *AVBPA 1997, LNCS, vol 1206*, pp. 1-9, 1997.
- [18] R. C. Gonzalez, and R. E. Woods, *Digital Image Processing*, 3rd ed., Pearson Hall, 2008.
- [19] L. Xu, E. Oja, and P. Kultanan, "A new curve detection method: Randomized Hough transform (RHT)," *Pattern Recognition Letters*, vol. 11, iss. 5, pp. 331-38, 1990.
- [20] A. A. Rad, K. Faez, and N. Qaragozlou, "Fast Circle Detection Using Gradient Pair Vectors," *VIIIth Digital Image Computing: Techniques and Applications*, pp. 879-887, December 2003.
- [21] M. Boyd, D. Carmaciu, F. Giannaros, P. Thomas, and S. William, "MSc Computing Science Group Project Iris Recognition," *Imperial College, London*, 2010.
- [22] T. J. Atherton, and D. J. Kerbyson, "Size invariant circle detection," *Image and Vision Computing*, vol. 17, pp. 795-803, 1999.
- [23] J. Gil, and Y. Rubio, "A new method for iris pupil contour delimitation and its application in iris texture parameter estimation," *CIARP 2005, LNCS, vol. 3773*, pp. 631 – 641, 2005.
- [24] A. Ross, and S. Shah, "Segmenting non ideal iris using geodesic active contours," *Biomet. Symp.: Special Session on Research at the Biometric Consortium Conference, Baltimore, USA*, pp. 1-6, 2006.
- [25] F. Silva-Mata, E. G. Llano, E. M. Alvarez Morales, and J. L. Gil Rodríguez, "A fast adaboosting based method for iris and pupil contour detection," *CIARP 2006, LNCS vol. 4225*, pp. 127 – 136, 2006.
- [26] R. Tang, J. Han, and X. Zhang, "Efficient iris segmentation method with support vector domain description," *Optica Applicata*, vol 39, no. 2, pp. 365-374, 2009.
- [27] J. Cui, Y. Wang, T. Tan, L. Ma, and Z. Sun, "A fast and robust iris localization method based on texture segmentation," *SPIE*, vol. 5404, pp. 401-408, 2004.
- [28] S. Kooshkestani, M. Pooyan, and H. Sadjedi, "A new method for iris recognition system based on fast pupil localization," *ICCSA 2008, Part I, LNCS, vol. 5072*, pp. 555–564, 2008.
- [29] N. Otero-Mateo, M. Vega-Rodríguez, J. A. Gomez-Pulido and J. M. Sánchez-Pérez, "A fast and robust iris segmentation method," *IbPRIA 2007, Part II, LNCS, vol. 4478*, pp. 162–169, 2007.
- [30] O. R. Yusupov, "A method of allocating reference contours for the purpose of recognition of the iris," *Scientific–technical of FerPI*, vol. 21, no. 2, pp. 13-18, 2017
- [31] CASIA, Iris image database, Ver. 4., Available at: <http://www.cbsr.ia.ac.cn/IrisDatabase.htm>

Processing An Effective Method For Clock Tree

Synthesis

Narek Avdalyan
CSD Nitro
Mentor A Siemens Business
Yerevan, Armenia
Narek_Avdalyan@mentor.com

Kamo Petrosyan
SG AMS
Synopsys Armenia
Yerevan, Armenia
Kamo.Petrosyan@synopsys.com

Abstract—The following article represents the method of clock tree synthesis (CTS). The method gives the advantage to synthesis clock tree with small clock skew and a smaller difference in the time of data and synchro signal spreading. The following method is applicable for each modern software of place and route (PR).

Keywords—Clock Tree Synthesis (CTS), Clock Skew, Data Required Time, Data Arrival Time

I. INTRODUCTION

The synthesis of a synchronous-signal tree has a direct impact on the speed of digital circuits. Below is the phrase (1) characterizes P calculation of the synchronized signal of the digital system.

$$P = s + d_{\max} + P_0 \quad (1)$$

Where s is the clock skew, d_{\max} is the worst way for the speeding, P_0 characterizes the constant of time skews so that we can calculate the correctness of period of the digital system. The skew of the synchro signal characterizes the maximum skew from the source of the synchronization signal to the entry of trigger synchro signal. The skew of the synchro signal is counted both on the ascending front and on the descending front, with corresponding triggers [2,3]. P_0 is a constant, which includes the time and the distribution of the data and the skews of other time dimensions. It is clear from the expression (1) that in order to minimize the P period of the synchro signal, it is necessary to reduce the skew of the synchro signal and at the same time the worst way of data distribution- d_{\max} . Since in submicron technology the interruptions conditioned by interlinks are gradually given more importance to, which means that synchro signal skew is in the greater dominance upon speed in the digital integral circuit (IC). Thus, the decrease of synchro signal skews is even more important in sequential digital IC. Consequently, the bigger the increasing number of sequential numeric ICs becomes the more complicated becomes the synthesis of synch signal spreading tree. Usually, the synchronization signal is distributed in a wide range of fan-out ways and must have access to all functional blocks.

II. THE METHODS OF THE SYNTHESIS OF SYNCH SIGNAL TREE

There are type ways of constructing synch signal trees-automatically and manually differentiated. Manually constructing a synchronized tree allows the designer to manage the number of synchro spreading levels, buffer count and buffer type at each level. Below we present the simplest example of building a synchronization tree manually, for

which you need to set the synchronization input PIN name (-root [get_port CLK]), the number of levels (-max_level_num 3), the buffer type (-use_lib_cells {CLKBUF16, CLKBUF8, CLKBUF4}) will be used during the synthesis of a synchronization tree. Obviously, in case of manual synthesis of a CTS, all timing parameters of the are evaluated by the designer as a result of which the choice of the number of levels or the number of buffer names is determined. Thus, manual synthesis of a synchro signal tree is a very complex and labor-intensive process, which is obviously not applicable in modern designs if we take into consideration the volume of modern projects and the complexity of the synchro signal tree. The given example is for smaller projects or for smaller pieces of large projects. Below is a simplest example of the synthesis of the tree for synchronous trees and the schematic appearance of the tree constructed in Fig. 1.

```
create_cts_spec -name manually_clk_functional_mode \  
-root [get_port CLK] \  
-max_level_num 3 \  
-use_lib_cells {CLKBUF16, CLKBUF8, CLKBUF4}
```

Fig. 1 represents the following: we have a sync sign, the name of which is CLK, which should distribute the synchronization signal in the n -number of triggers. It is necessary to divide the synchronization signal into three levels (hierarchies) 1st, 2nd and 3rd. In each level, use different buffers with corresponding X16, X8 and X4.

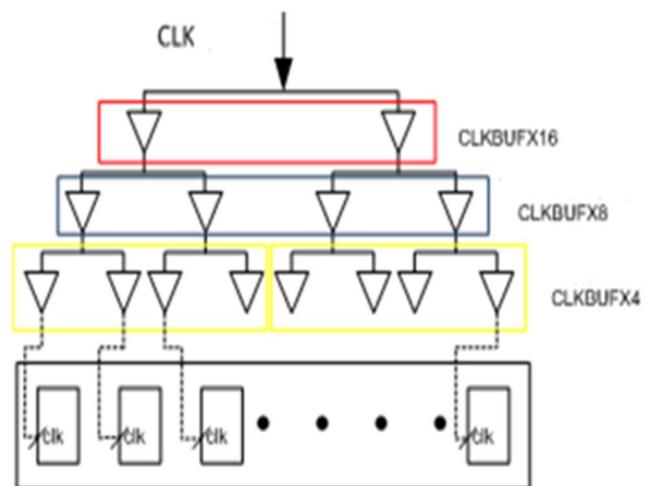


Fig. 1. Schematic appearance of the manually constructed synchro signal tree

The software tool that has been developed by us at the time of building an automatic synchronization tree. Deployment

and scheduling (e.g. Nitro-SoC using the following commands: `report_cts`, `report_timing`, `report_cts`) [2] automatically retrieves all time-related settings for the synchronization tree from the software environment. For the method of constructing an automatic synchronization tree developed by us, the following parameters are required for the maximum, minimal delay (`-max_delay 450p`, `-min_delay 400p`), maximum duration (increasing, decreasing) duration (`-max_transition 50p`) for interconnection and cell (`-max_leaf_transition 50p`) maximum capacity (`-max_wire_capacitance 400.0f`) for interconnection and cell (`-max_capacitance 200.0f`), maximum output branch (`-max_fanout 10`), maximal skew of synchronization signal (`-max_skew 30p`) cell for interconnect (`-max_net_skew 50p`). Based on the data received and processed, synchro signal tree topology is constructed and is balanced, including the number and size of the buffers calculated and imported by the method.

Regardless of what kind of methodology is used for the synthesis of a CTS, the main goal is to minimize the time synchro signal distribution and to provide a skew as little as possible the synchro signal (distraction). Let's consider the synchro skew and the time based on the simpler example of synchro signal spreading Fig. 2.

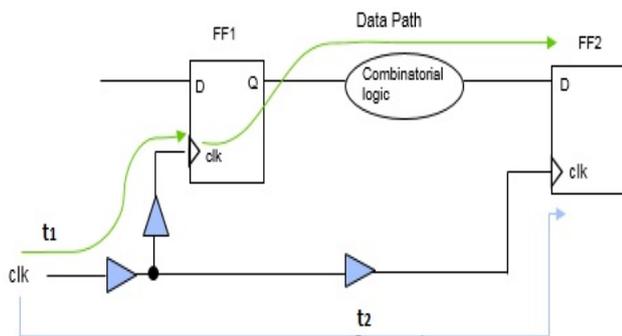


Fig. 2. The synchrotron skew between two triggers

There are two FG1 and FF2 triggers that are linked with Combinatorial Logic and Sync with `clk` [1]. The skew of synchro signal will be represented with the following formula.

$$\delta = t_2 - t_1 \quad (2)$$

Where δ is skew of synchro signal between two triggers, t_2 , t_1 are respectively time for synchro signal spreading.

Below is the data required for the synthesis of an automatic CTS. Fig. 3 is presented in the real design of the CTS.

```
create_cts_spec -name auto_clk_functional_mode \
-root [get_port CLK] \
-max_transition 50p \
-max_leaf_transition 50p \
-max_capacitance 200.0f \
-max_wire_capacitance 400.0f \
-max_fanout 10 \
-max_skew 30p \
-max_net_skew 50p \
```

```
-min_delay 400p \
-max_delay 450p \
-opt_clock_gates {move size} \
-cell_name_prefix CTS_clk_cell \
-port_name_prefix CTS_clk_port \
-use_inverters true \
-trace true \
-balance_roots true
```



Fig. 3. The appearance of CTS

As can be seen from the example, all the parameters needed for the synthesis of an automatic CTS are represented with their respective values. It is also important to group synchro signals that is based on skews of synchro signals, and the synchro signals that are as closest in their values as possible are grouped together. For a more detailed description of the method, we will unveil the nature of our method as follows. Follow the steps below.

III. CONSIDERING THE METHOD SUGGESTED BY US

Obviously, allocation and software tools for the synthesis of the synchronous signal tree use quite complex and different algorithms. The purpose of our method is to minimize the skew of the synch signal and reduce the speed of synchro signal in CTS. Thus, we need to pre-evaluate the synthesis of CTS, based on which we should apply our method in future. First, it is necessary to find the greatest skew with synchro triggers and obviously, allocation and software tools for the synthesis of the synchronous signal tree use quite complex and different algorithms. The purpose of our method is to minimize the skew of the synch signal and reduce the speed of synchro signal in CTS. Thus, we need to pre-evaluate the

synthesis of CTS, based on which we should apply our method in future. First, it is necessary to find the greatest skew with synchro triggers and the longest time of synchro signal movement. Next we have to look for the next greatest values until we find them all. Afterwards, we should group by descending order of the two parameters listed above. We also need to consider the distraction of the two parameters taking into consideration the values we receive. The greater the distraction is, the worse synchro signal tree is constructed. We calculate the distraction in the following way: We use three sample values 9 ps, 10 ps, and 11 ps. First, the average of these samples is calculated:

$$(9 + 10 + 11) / 3 = 10 \quad (3)$$

The average is 10ps. Next, the squared deviation is calculated:

$$(9 - 10)^2 + (10 - 10)^2 + (11 - 10)^2 = 2 \quad (4)$$

The squared deviation is 2 ps. Finally, jitter is calculated from the square root of the average squared deviation.

$$\sqrt{2 / (3 - 1)} = 1 \quad (5)$$

Jitter is 1ps

So let's consider the method suggested by us which consists of two phases. Below we present the sequence of steps:

First phase

- It is necessary to group synchronous signal skews by ascending order.
- It is necessary to evaluate the skews of synchronization signals and distraction times.
- On the basis of the values obtained, it is necessary to estimate the number of hierarchies of the synchronization signal, taking into account the timing parameters of the buffers and the inverters in the standard library.
- The choice of the correct quantity of hierarchies is very important because the smallest possible value of the skews depends on it.
- It is necessary to set a threshold value to which the skew of synchro signals should strive after choosing the right quantity of hierarchies.
- In other words, the hierarchies of synchro signals will increase so much that we can reach our threshold value.
- After determining the number of hierarchies's, synchro signals are again recalculated and classified in ascending order.

Second phase

- The CTS that has the biggest skew from the classified ranges is considered and we try to reduce the deviation by 2 stages by changing the size of the existing buffer.
- In case of failing the 1st step, we are going to optimize the data transmission path.

- It is necessary to give valuation after each step in order to avoid problems of optimization in our suggested method.
- The described method is automated but if required it allows the designer to interfere in the method.

Let's consider the above mentioned method in practice. Fig. 4 depicts some part of CTS, where synchro signal is underlined. Fig. 5 clearly envisages that synchro signal comes out of the buffer (is marked with red circle) and spread between 4 triggers(is marked in yellow circle). As a result of the analysis it becomes clear that, in this case, we have the following skews of the synchro signal Table I.

TABLE I. Skews Of The Synchro Signals (Before Method Implication)

The Number Of The Trigger	Synch Signal Skew (ps.)
1	25.4
2	26.9
3	24.1
4	19.4

As a result of method implication we receive the following results Table II:

TABLE II. Skews Of The Synchro Signals (After Method Implication)

The Number Of The Trigger	Synch Signal Skew (ps.)
1	16.7
2	17.4
3	18.3
4	13.8

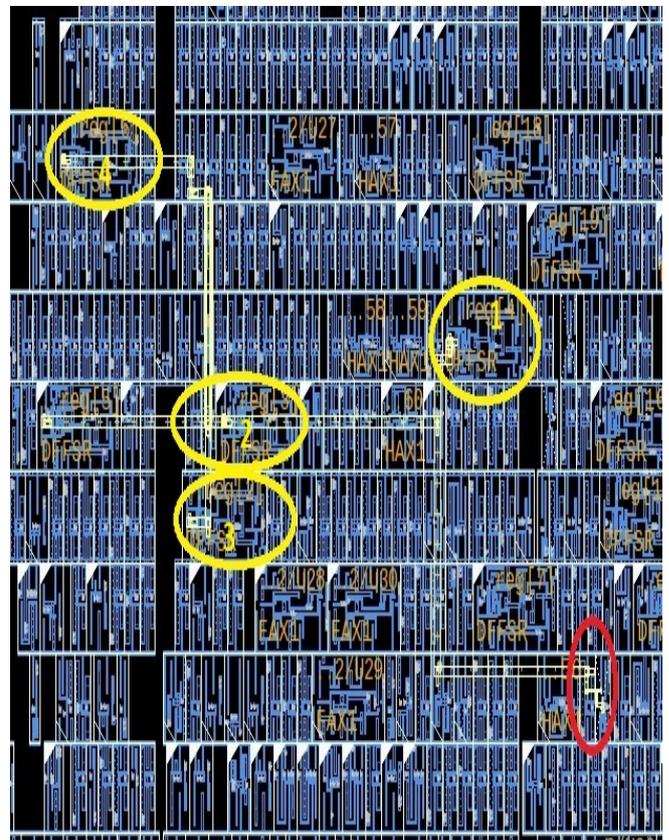


Fig. 4. Part of CTS

Fig. 5. presents the most remote placed triggers of our project(they are marked with red). It is obvious that from the angle of the CTS synthesis it would be quite difficult to minimize the time of synchro signal spreading. In this context, it is necessary to consider the time of movement of the data and try to maximize it so that we can neutralize the amount of time it takes to distribute the synchronized signal at a great distance. At the same time, we can also try to add additional buffers on the way of the synchro signal so that we do not significantly increase the time of data motion. In this case, the application of the method developed by us gave the following result:

Before the application of the method the difference between the synchro signal and the time of propagation was 356 ps.

After the application of the method it became 48ps.

Thus, the project is comprised of 1213 triggers, which is a memory control unit (SRAM) block with static self-willed permissions. Thus, it is necessary to construct a CTS so that the difference between the clock skew and the spread of the data and the synchro signal (slack) satisfies the thresholds we propose. As much we can design projects with smaller threshold as faster the designed project will be. The proposed method can be used to build a CTS in any of the placement and routing tools (IC Compiler, Nitro-SoC, Innovus). This method not only reduces the synchro signal the skew, but also reduces the difference between the time of spreading the data and the synchro signal.

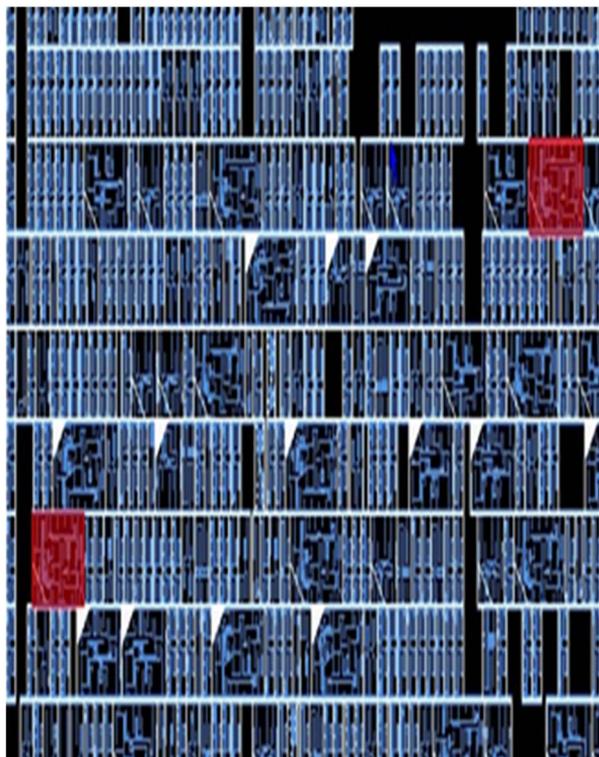


Fig. 5. The most distant-placed triggers

Obviously, the method developed by us allows to reduce the skew of the synchro signal, the difference of spreading time between data and synchro signal by dividing the additional levels and classifying it by time. By adding additional buffers on these classified ways, we increase the area of our project. It is also important to note when reducing synchro signal skew it is advisable to make use of the method

suggested by us. If necessary, it is possible to add double inverter instead of a buffer.

Below Table III represents the skew of the synchro signal before the application of our method, it contains 11 levels. Table IV represents synchro signal skew after the application of the method when the levels have been added reaching 20 as a result of buffers added by us.

TABLE III. Total Skews Of The Synchro Signals (Before Method Implication)

The Name Of Synchro Signal	Skew (ps.)
CLK_1	25.4
CLK_2	26.9
CLK_3	24.1
CLK_4	19.4
CLK_5	18.8
CLK_6	18.4
CLK_7	16.2
CLK_8	28.3
CLK_9	19.7
CLK_10	18.1
CLK_11	18.5

TABLE IV. Total Skews Of The Synchro Signals (After Method Implication)

The Name Of Synchro Signal	Skew (ps.)
CLK_1	16.7
CLK_2	17.4
CLK_3	18.3
CLK_4	13.8
CLK_5	11.3
CLK_6	11.2
CLK_7	11.5
CLK_8	20.3
CLK_9	15.8
CLK_10	12.2
CLK_11	19.8
CLK_12	20.2
CLK_13	19.7
CLK_14	18.8
CLK_15	14.1
CLK_16	15.1
CLK_17	17.9
CLK_18	15.3
CLK_19	16.4
CLK_20	16.1

Table V represents the difference of spreading time between synchro signal and data before and after the application of the method.

Table V. The difference of spreading time between synchro signal and data before and after the method.

Before The App. Of The Method (ps.)	After The App. Of The Method (ps.)
356	48

IV. CONCLUSION

The proposed project had an average of 21.2 ps. (Jitter 4.12 ps.) synchro signal skew and 11 levels before using our reduction method. As a result of our method, the number of levels has increased to 20, and the synchro signal skew is 16.1 ps. (Jitter 3.01 ps.), and the area has increased to 15.45%, the overall project performance has increased to 4.78%. The difference of spreading time between the data and the synchro signal was at least improved 7 times.

ACKNOWLEDGMENT

This paper was supported by Mentor A Siemens Business. We thank our colleagues from CDS Nitro division who provided insight and expertise that greatly assisted the research. We thank Irina Dumanyan (Mentor A Siemens Business, Armenia Site Manager) for initiating and supporting this work, Armen Ketikyan (Department Manager CSD Nitro), Vahe Arakelyan (Group Manager, CDS Nitro).

REFERENCES

- [1] Chentouf Mohamed, Alaoui Ismaili Zine El Abidine "Physical Design Automation of Complex ASICs" IJCSI International Journal of Computer Science Issues, Volume 15, Issue 1, January 2018 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784. www.IJCSI.org <https://doi.org/10.20943/01201801.1424>
- [2] Nitro-SoC User's Manual, Software Version 2017.2, February 2018.
- [3] Encounter User Guide, Product Version 5.2.1, February 2006.
- [4] Tsay, R.-S. "Exact zero skew," Computer-Aided Design, 1991. ICCAD-91. Digest of Technical papers., 1991 IEEE International Conference, pp. 336-339, 1991.
- [5] Madhav P. Desai, Radenko Cvijetic, and James Jensen. "Sizing of clock distribution networks for high performance cpu chips." In Proceedings of the 33rd annual conference on Design automation, pp. 389-394, 1996.
- [6] Simon Tam, Stefan Rusu, Utpal Nagarji Desai, Robert Kim, Ji Zhang, and Ian Young. Clock generation and distribution for the first IA-64 microprocessor. IEEE Journal of Solid-State Circuits, 35(11):1545-1552, Nov 2000

Development of Automation Systems at Transport Objects of MegaCity

Andrei Belyi
Bridges Department
Emperor Alexander I St.
Petersburg State Transport
University
Saint-Petersburg, Russia
andbelyi@mail.ru

Dmitrii Shestovitskii
Bridges Department
Emperor Alexander I St.
Petersburg State Transport
University
Saint-Petersburg, Russia
shestovitsky@mail.ru

Valerii Myachin
General Director
STC "Complex monitoring
systems" LLC
Saint-Petersburg, Russia
vnmyachin@yandex.ru

Dmitrii Sedykh
Department of Automation
and Remote Control on
Railways
Emperor Alexander I St.
Petersburg State Transport
University
Saint-Petersburg, Russia
sedyhdmiriy@gmail.com

ABSTRACT. In article it is executed comprehensive and systematic the analysis of objects of transport infrastructure of the megacity. The research of emergence and functioning of systems of automation on constructions of St. Petersburg, their evolutionary development and transformation to the single systems of tool monitoring is executed. For the first time use of monitors for the park of artificial constructions is proved. On the basis of detailed inspection of city facilities (in total more than 700 units) criteria were formulated and groups of transport objects which allowed to allocate about 100 constructions are created, without fail due to be equipped with monitors

Keywords—monitoring, automation systems, transport objects, diagnostics, technical condition

I. INTRODUCTION

Tool monitoring systems (or structural health monitoring systems) have long been firmly established in engineering life. Their widespread use is due to the significant progress in the field of information technology observed throughout the world over the past 30-40 years.

Monitoring as a research tool first began to be used in the 70s of the twentieth century. Initially, monitoring meant an environmental observation system that controls the processes of interaction between nature and man. In papers and scientific studies on nature-oriented and environmental matters, monitoring quite quickly becomes one of the most widely used concepts [1]. Currently, various environmental monitoring systems have been developed and are functioning, which determine the organization of constant observations in space and time over the man-made changes in natural environment and control of its state during various types of economic activity [2].

The general principles of environmental monitoring served, in particular, as the basis for the creation of engineering monitoring as a new direction in the field of organizing the operation of complex building structures, including artificial road structures [3].

Ensuring sustainable operation requires constant surveillance over the appearance of certain defects and damage of the object elements, and forecasting their possible development before they turn into defects and damage that threaten structure reliability and durability.

It should be noted here that there are two main types of monitoring: one is applied during the construction process and another is applied during the operational period; both of them solve essentially different tasks.

While the first type controls the stress-strain state of a structure during construction, when non-design forces occur in structures, the second one is intended mainly to monitor the technical condition of the structure affected by negative factors and influences during its existence for quite a long time.

In our case, using the term *monitoring*, we mean precisely monitoring during the operational period.

Let us give examples of existing innovation approaches to managing the technical condition of artificial structures and monitoring systems both in Russia (in particular, in St. Petersburg) and abroad.

The world experience in implementing monitoring systems is rather extensive [4, 5]. In domestic practice in recent years, many objects of transport infrastructure have been equipped with similar systems [1, 5]. At the same time, a number of publications relating to the monitoring of urban structures can be singled out [1-5].

St. Petersburg, being a kind of museum of bridges, as the owner of artificial structures park, whose architectural and technical features are recognized all over the world [1-5], could not stand aside.

In 2017-2018, Mostotrest St. Petersburg State Budgetary Institution, the oldest operating organization in the field of management of the technical condition of transport facilities, assigned a task to develop *The concept for monitoring over artificial road structures in St. Petersburg using automated technologies with the subsequent development of working documentation for the automated monitoring system of the Alexander Nevsky Bridge* (hereinafter referred to as the Concept of Monitoring). In order to provide services timely and with correspondence quality in accordance with the state contract, experts in the industry who are co-authors of this article were involved in the work as experts and co-developers.

The purpose of the article is to describe its main provisions, since in fact it was a research work that allows highlighting some issues in the field of monitoring over artificial structures and the process of automating the maintenance of St. Petersburg bridge park objects.

In the course of Monitoring Concept development, a number of activities were carried out aimed at analyzing the development of existing automation systems for facilities in St. Petersburg, as well as their current status and future development.

The objectives of the article are city road artificial constructions and their subsystems.

II. SYSTEM DEVELOPMENT AND OBJECT ANALYSIS

It is widely known that St. Petersburg has unique artificial structures, characterized by both aesthetic and technical features that are recognized throughout the world. This park of bridges includes a wide variety of structures of multifarious materials and static schemes, whose service life is comparable to the age of the city itself.

Despite of a great variety of unique artificial structures, first of all, they are the transport structures which main purpose is to ensure constant, safe and uninterrupted movement on the highways that they connect.

Accordingly, it is necessary to operate and maintain not every bridge separately, but the entire bridge park as a whole; not to save each separate bridge, but classify it as a unit of the bridge park.

Moreover, if during the operation of individual structures each of them should serve as long as possible, then the bridge park as a whole should be operated continuously [6].

There is no need to support bridges for 120-150 years. Such durability can be technically ensured, but it is not economically feasible.

It is much more profitable to ensure safe operation and loading pass at the proper level, i.e. the most important function of the structure during the so-called optimal period.

According to calculations, this period does not exceed 65-70 years under the conditions of St. Petersburg (for example, reinforced concrete bridges).

In addition to the use of modern equipment when maintaining bridges [7], which allows maintaining reliability and functionality required levels, continuous improvement of the existing operation system is also required, which in turn requires increased attention and the use of additional resources (both material and intellectual).

Respectively, it is necessary to use modern methods and means of monitoring the technical condition of structures subject to the term *innovative*.

Let us give some examples.



Fig. 1. The system of radar monitoring, security and video surveillance in the premises of the Troitsky Bridge

In order to increase the level of buildings protection, complex bridges and security systems have been installed on

movable bridges since the 1990s of the twentieth century (Fig. 1), which allow monitoring the presence of vehicles and pedestrians before raising the bridges from the control panel and the guard station, as well as warning unauthorized penetration to bridges and service premises, and also damage to property and equipment.

In addition, the installed radar system allows the dispatcher to control ships passage along the Neva River at night along the sections of reconnected spans. It is extremely necessary for use in emergencies cases with dockings impact with the bridge piers (Fig. 2)



Fig. 2. Vessel after impact with the Troitsky Bridge support

The subsystem of radar and visual monitoring of ships passage along the Neva River fairway allows determining craft parameters (dimensions, exact location geographical coordinates, movement speed and direction) in real time, and also displays craft current location in online mode on an electronic map with reference to geographical coordinates. Information is being recorded and archived, and a database of alarm events is maintained.

Currently, video surveillance and security systems are very widespread due to the relative ease of operation and significant advantages in maintaining the structure. This is especially concerns a matter of ensuring the safety of St. Petersburg treasury to which the artificial structures of the city belong together with the equipment installed.

The system of automated control over bridges raising applies with the following objects:

- The Volodarsky Bridge;
- The Alexander Nevsky Bridge;
- The Bolsheokhtinsky Bridge;
- The Liteyny Bridge;
- The Troitsky Bridge;
- The Dvortsovy Bridge;
- The Blagoveshchensky Bridge;
- The Birzhevoy Bridge;
- The Tuchkov Bridge.

The system of automated control over the bridges raising provides dispatch control of power supply and monitoring over the process equipment operation supporting bridges raising, as well as the transmission of video information from cameras of bridges technological video surveillance. On some bridges, the system provides control and transmission of the alarms from the security, alarm and fire alarm systems and automatic fire extinguishing to the control room.

The system's dispatch center is in the administrative building of Mostotrest St. Petersburg State Budgetary Institution located at 42, Industrialny Avenue (Fig. 3)



Fig. 3. The dispatch center of Mostotrest St. Petersburg State Budgetary Institution located at 42, Industrialny Avenue

Access to the system data is provided through the LAN (local area network) of Mostotrest St. Petersburg State Budgetary Institution, including remote access from the premises of operation section of movable bridges located at 4, Orlovsky Lane.

The automated traffic safety system in tunnels covers the following objects:

- The left-bank transport tunnel at the Liteyny Bridge;
- The right-bank transport tunnel at the Liteyny Bridge;
- The tunnel of transportation hub of the right-bank exit from the Liteyny Bridge;
- The transport tunnel at the Grenadersky Bridge;
- The transport tunnel under the Pobedy Square;
- The pumping station of overpasses at the intersection of road to the airport with Pulkovskoye Highway;
- The transport tunnel on Pulkovskoye Highway (turn to Pushkin Town);
- The overpass of tunnel type (the tunnel on Pirogovskaya Embankment);
- The pumping station of railway junction on Ligovsky Avenue;
- The underground pedestrian crossing on the Mitrofanyevskoye Highway;
- The underground pedestrian crossing under the Primorsky Avenue (at the 3rd Yelagin Bridge).

The automated traffic safety system in tunnels provides dispatch control of power supply and monitoring over the operation of technological equipment of sewage pumping stations, ventilation and outdoor lighting of transport tunnels and underground pedestrian crossings. The system provides the ability to remotely control devices for automatic input of backup power (ATS) and pumping equipment. At some sites, sensors of security, alarm and fire alarms are included in the system. Recently equipped facilities (the right-bank transport tunnel at the Liteyny Bridge, the tunnel of transportation hub of the right-bank exit from the Liteyny Bridge, the tunnel on Pirogovskaya Embankment) provide video data transmission from process video surveillance cameras.

The system's dispatch center is combined with the *Moveable Bridges of St. Petersburg System's Dispatcher Center* and is located in the administrative building of Mostotrest St. Petersburg State Budgetary Institution located at 42, Industrialny Avenue.

The dispatching system of elevating equipment for pedestrian crossings covers the following objects:

- The elevated pedestrian crossing No. 1 opposite Oskalenko Street;
- The elevated pedestrian crossing No. 2 on Primorsky Avenue;
- The elevated pedestrian crossing in the alignment of Staroderevenskaya Street;
- The elevated pedestrian crossing at Savushkina Street;
- The elevated pedestrian crossing at Yakhtennaya Street.

The dispatching system of elevating equipment for pedestrian crossings provides dispatcher control of lifting equipment and control access to lifting platforms for low-mobility groups of people at elevated pedestrian crossings, including video transmission from surveillance cameras and two-way voice communication between the user of lifting equipment and the dispatcher.

The system's control room is equipped in an underground pedestrian crossing near Primorsky Avenue (at the 3rd Yelagin Bridge). The system's information is duplicated and stored in the control room, combined with the *Moveable Bridges of St. Petersburg System's Dispatcher Center* and located in the administrative building of Mostotrest St. Petersburg State Budgetary Institution located at the following address: 42, Industrialny Avenue.

The underwater transport tunnel under the Morskoy Channel on Kanonersky Island can be one more example, where, along with a video surveillance and security system, a system for recording data on the technical condition and performance of equipment and electrical networks in a tunnel is installed (Fig. 4).

The autonomous monitoring and control system for engineering systems of the Kanonersky transport tunnel provides work monitoring and remote control of the tunnel engineering systems (pumping stations, ventilation systems, tunnel lighting), as well as technological video surveillance.



Fig. 4. CCTV and security systems in the Kanonersky transport tunnel

The system's control room is located in the technical building of the Kanonersky tunnel.

The system of operational monitoring of the overpass through the railway tracks of St. Petersburg-Sortirovochny-Moskovsky station in the alignment of Alexandrovskaya Ferma Avenue provides monitoring over the overpass building structures and technological video surveillance (Fig. 5)

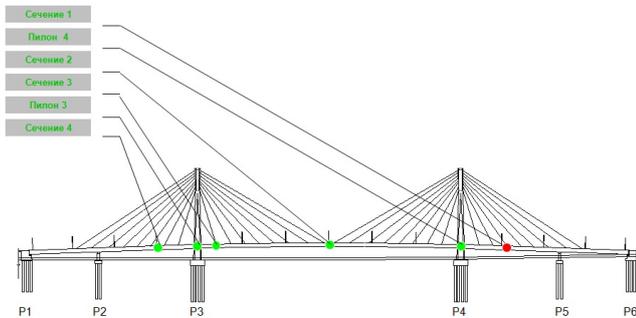


Fig. 5. Monitoring over the state of the overpass constructions in the alignment of Aleksandrovskaya Ferma Avenue

It is obvious that such systems are of a great benefit in buildings maintenance, because they seriously increase the efficiency of the operation process by reducing labor costs (as well as material resources) to protect the structure. In addition, as it was already mentioned, the process of investigating accidents on artificial structures is greatly simplified.

However, these systems are already far from new ones. Their use began in the middle 90s of the twentieth century. Nevertheless, their further improvement is possible, and is constantly being developed.

Currently, there are more than 700 artificial structures in St. Petersburg in the technical maintenance of Mostotrest St. Petersburg State Budgetary Institution.

Due to the fact that a significant part of the facilities has been constructed according to individual projects, and are the structures of *complex* design, and in some cases even attributed to particularly dangerous, unique and technically complex (in accordance with regulations of the Russian Federation), there is an obvious need for special approaches to the management of their technical condition (maintenance).

Operational control over the technical condition of buildings and structures is carried out during the period of operation of such buildings and structures by conducting periodic inspections, checking (or) monitoring of the condition of bases, building structures, engineering systems and engineering networks in order to assess the state of structural and other characteristics of reliability and safety of buildings and structures (in accordance with regulations of the Russian Federation).

Also, both in Russian and international standards, it is explicitly stated that in applicable case for the purpose of evaluating the actual operation of bridge structures, monitoring of the stress-strain state of bridges, i.e. a system of long-term monitoring of their condition and behavior in

the process of construction (reconstruction) and operation should be provided in projects.

Monitoring should be organized in the following cases:

- during the construction and operation of *large*¹ and *complex* bridges;
- for metal and reinforced concrete structures wherein their additional prestressing is applied (force control);
- for bridges with externally statically indefinable structures, wherein additional forces, deformations and sediments are possible due to geological, hydrological, landslide and seismic phenomena;
- for reinforced concrete structures wherein large uncertainty of long-term processes associated with creep, shrinkage and temperature deformations (concrete different ages, a combination of prefabricated and monolithic structures, etc.) is possible.

In accordance with the Russian legislation, the objects of control, threats of accidents, and emergencies, should be subsystems of life support and safety:

- heat supply;
- ventilation and air conditioning;
- water supply and sewerage;
- power supply;
- gas supply;
- engineering and technical complex of the facility's fire safety;
- lift equipment;
- communication system and alerts;
- alarm systems, video surveillance, access control and monitoring, as well as inspection equipment;
- systems for detecting increased levels of radiation, emergency chemical substances, biohazardous substances, significant concentrations of toxic and explosive concentrations of gas-air mixtures, etc.

Objects to control the threat of accidents and emergencies should be technological systems, as well as bases, engineering structures of buildings and facilities; engineering protection facilities, areas of possible mudflows, landslides, and avalanches in the area of the facility operation.

Thus, when analyzing potential monitoring objects, several characteristic groups of objects were identified, fully or partially related to objects equipped with an automated monitoring system.

The first group of objects consists of artificial road structures attributed to large bridges in accordance with the standards. Mostly, moveable and non-moveable bridges, as well as city overpasses are included in this group.

¹ Note: Here and elsewhere it is assumed that small bridges are up to 25 m long inclusive; medium bridges are over 25 m long and up to 100 m inclusive; large bridges are over 100 m long. Highway (including city) bridges which are less than 100 m long, but with spans of over 60 m are also belong to large bridges.

The second group of objects consists of transport tunnels. The objects are of complex design; the safety of their operation using monitoring systems is significantly increased. The risks analysis of emergencies development for these structures and similar objects [8, 9] using existing standards (GOST R 51901-2002. Reliability management. Risk analysis of technological systems) and scientific research [2, 3] confirms this need.

The third group of objects consists of pedestrian tunnels, pumping stations, lifting equipment and other engineering networks at St. Petersburg transport infrastructure objects. Mostly, communications and objects' engineering networks (the so-called engineering systems monitoring process) are the elements under control.

The fourth group consists of other objects, for example, structures that have a relatively small length, but which are complex in design, or that have emergencies or similar situations during their life cycle. An express analysis of emergencies risks at such facilities confirms the need for automated monitoring systems to ensure the safety of their operation. These are dams, bridges less than 100 m, and complex structures.

Typical architecture of the monitoring system is the following – fig. 6.

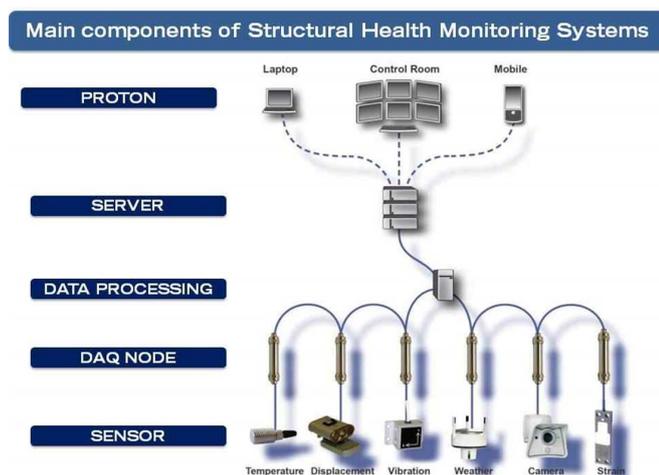


Fig. 6. Typical architecture of the monitoring system

Moreover, some systems allow monitoring moving objects. This can be realized by special additional sensors in accordance with CCTV components. For instance let us show an equipment from already pointed overpass constructions in the alignment of Aleksandrovskaia Ferma Avenue – fig. 7.

The overpass is instrumented in four various sections and also at the level of two poles and cross beams (struts). Besides temperature sensors and strain gauges, the control system of loadings with possibility of video fixing and registration of the uncontrollable superheavy vehicles which are carrying out unauthorized driving through the overpass acts on the overpass. For this purpose on object the video cameras allowing when receiving alarm signals from sensors of monitoring are installed to register images for 5 with before inclusion of alarm signal and 10 from later its inclusion.



Fig. 7. Monitoring of moving objects: left photos – cameras on the pylon according with strain gauges; right photos – fixed vehicles.

The control system of loadings gives the chance to estimate the strain (stressed-deformed) level which is tested by designs when passing heavy-load cars, and to establish threshold on reaching which it is necessary to survey design. It is also possible to count cycles of passing of motor transport and to estimate the level of fatigue of object, i.e. to analyse change of the main mechanical and physical properties of designs under the influence of the cyclic tension and deformations.

Before installation of system risk analysis, the heavy-load cars connected with passing on object which has revealed the specific zones serving as the indicator when tracking influence of this risk (fig. 7) has been made.

From the aforesaid, it can be seen that potential monitoring sites have some regularities in relation to the city's transport infrastructure:

- Bridges are located on the Neva River and the branches of its delta;
- There are various objects (bridges, transport tunnels, pedestrian crossings) along the northern shore of the Neva River delta (Sverdlovskaya Embankment, Pirogovskaya Embankment, Ushakovskaya Embankment, Primorsky Avenue, Savushkin Street, Primorskoye Highway);
- Increased density of placement of potential monitoring objects is observed along the main radial (Ligovsky – Moskovsky Avenue – Pulkovskoye Highway, Vitebsky Avenue, Stachek Avenue – Marshal Zhukov Avenue – Tallinskoye Highway) and city ring roads (Obvodny Channel Embankment, Ivanovskaya Street – Slavy Avenue).

In total, more than 100 objects were included in Monitoring Concept.

CONCLUSION

So far, there have been isolated cases of the use of tool monitoring means in our country, and for the first time the authors have proposed a full-fledged concept covering all city road objects.

Together with the fact that St. Petersburg is a megacity with various man-made structures of absolutely all types (and many of them are monuments), tool monitoring of the structures technical condition is a complete and relevant mean aimed at improving the efficiency of artificial road structures. It is based on the physical laws and automation algorithms, jointly allowing putting modern monitoring systems into practice.

The presence of more than 700 artificial road structures in the city led to the fact that in the process of Monitoring Concept developing, a complete and reliable facilities analysis was needed along with the development of clear criteria for equipping them with monitoring tools. The authors solved this problem by forming four groups of objects. Otherwise, a situation of redundancy monitoring and a subsequent decrease in the system efficiency could occur.

The primary objects for Monitoring Concept implementation are large transport infrastructure facilities, such as transport tunnels and drawbridges.

The main novelty of the article is representing and justification of constructions types to be installed by structural health monitoring systems.

REFERENCES

- [1] S. Sumitro, M.L. Wang., "Structural Health Monitoring System Applications in Japan," In: Ansari F. (eds) *Sensing Issues in Civil Structural Health Monitoring*. Springer, Dordrecht, 2005. – pp. 495-504. https://doi.org/10.1007/1-4020-3661-2_49.
- [2] J. E. Andersen, M. Fustinoni, *Structural health monitoring systems*, Italy: L&S S.r.l. Servizi Grafici, 2006. – 126 p.
- [3] W. Rucker, F. Hille, R. Rohrmann, *Guideline for structural health monitoring*. Final report, SAMCO, Berlin, 2006. – 63 p.
- [4] H. Wenzel, *Health monitoring of bridges*, Chichester: John Wiley & Sons, 2009. – 621 p.
- [5] Efanov D., Osadchy G., Sedykh D. "Development of Rail Roads Health Monitoring Technology Regarding Stressing of Contact-Wire Catenary System", In *Proceedings of 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, Chelyabinsk, Russia, 19-20 May, 2016, doi: 10.1109/ICIEAM.2016.7911431
- [6] Efanov D., Osadchy G., Sedykh D., Pristensky D., Barch D. "Monitoring System of Vibration Impacts on the Structure of Overhead Catenary of High-Speed Railway Lines", In *Proceedings of 14th IEEE East-West Design & Test Symposium (EWDTS'2016)*, Yerevan, Armenia, October 14-17, 2016, pp. 201-208.
- [7] Efanov D., Pristensky D., Osadchy G., Razvitnov I., Sedykh D., Skurlov P. "New Technology in Sphere of Diagnostic Information Transfer within Monitoring System of Transportation and Industry", In *Proceedings of IEEE East-West Design & Test Symposium (EWDTS 2017) Proceedings*. 2017. pp. 231-236.
- [8] Efanov D., Sedykh D., Osadchy G., Barch D. "Permanent Monitoring of Railway Overhead Catenary Poles Inclination", In *proceedings of IEEE East-West Design & Test Symposium (EWDTS 2017) Proceedings*. 2017. pp. 163-167.
- [9] A.A. Belyi, E.S. Karapetov, Yu. S. Efimenko, "Structural health and geotechnical monitoring during transport objects construction and maintenance (Saint-Petersburg example)", *Procedia Engineering*. – 2017. – Vol. 189. pp. 145–151. PII:S1877-7058(17)32145-8 DOI:10.1016/j.proeng.2017.05.024.

Advanced Indication of the Self-Timed Circuits*

Yury Stepchenkov

*Institute of Informatics Problems
Federal Research Center "Computer
Science and Control" of the Russian
Academy of Sciences
Moscow, Russia
YStepchenkov@ipiran.ru*

Yury Shikunov

*Institute of Informatics Problems
Federal Research Center "Computer
Science and Control" of the Russian
Academy of Sciences
Moscow, Russia
yishikunov@gmail.com*

Yury Diachenko

*Institute of Informatics Problems
Federal Research Center "Computer
Science and Control" of the Russian
Academy of Sciences
Moscow, Russia
diaura@mail.ru*

Denis Diachenko

*Institute of Informatics Problems
Federal Research Center "Computer
Science and Control" of the Russian
Academy of Sciences
Moscow, Russia
diaden87@gmail.com*

Yury Rogdestvenski

*Institute of Informatics Problems
Federal Research Center "Computer
Science and Control" of the Russian
Academy of Sciences
Moscow, Russia
YRogdest@ipiran.ru*

Abstract—Paper discusses a problem of the CMOS self-timed circuits' indication. Large number of indicating signals in the multi-bit computational devices and registers requires an additional hardware and time for their combining and forming a single control signal that provides a request-acknowledge interaction between interconnected self-timed functional blocks. Indication subcircuit performs this. Multi-input hysteretic triggers allows for accelerating indication subcircuit by factor of 1.1 – 1.6 and reducing its complexity in several times in comparison to standard implementation basis on static and semi-static Muller's elements. A penalty for this is some short-circuit current in the worst case.

Keywords—self-timed, indication, C-element, hysteretic trigger, performance, complexity

I. INTRODUCTION

Theoretically self-timed (ST) circuits provide best performance in any particular ambient conditions because they are free of any given external clock. They use a request-acknowledge interaction between source of a processed digital data and its receiver. Due to this performance of the ST circuits is determined only by the real cell delays in the current operating conditions.

Unlike ST circuits, synchronous circuits operate under strict control of an external clock. Therefore, if a clock source does not adapt to changing ambient conditions, synchronous circuits are forced to focus on the "worst case": minimum supply voltage, maximum permissible ambient temperature, "slow" transistors etc. As a result, in some applications, the ST circuits are faster than their synchronous counterparts despite their hardware redundancy.

The main factors limiting performance of the ST circuits are as follows:

- Diphas work discipline.
- Presence of an indication subcircuit.

There are two phases in any ST circuit operation: work phase implementing an input data processing algorithm and spacer, in which ST circuit prepares for next work phase. Spacer is necessary to separate adjacent work phases, but it adds nonproductive delay to a total work cycle of ST circuit.

Indication subcircuit is an integral part of the ST circuits. It provides completion detection at each phase and controls an interaction between ST functional blocks. Indication subcircuit combines all internal indication signals into one phase signal. It is a control signal for ST circuits that are the drivers and receivers regarding this ST circuit. ST circuit is considered to be switched to the next phase only when both algorithmic part of the circuit, and indication subcircuit have switched to this phase. At that, all circuit elements must complete their switch in this phase. Therefore, to confirm the end of switching ST circuit, one needs to indicate outputs of all circuit components and combine them into a single indication output.

The higher performance of the ST circuits compared to their synchronous counterparts is showed obviously in relatively simple circuits with a small capacity. Here indication subcircuit works in the "background" mode and its contribution to the circuit delay is negligible. In multi-bit ST circuits a large number of internal indication signals leads to "swelling" indication subcircuit and to increasing its contribution to the digital data processing tract delay.

Therefore, the development of the components accelerating completion detection of the ST circuits is an urgent task. This paper analyses the indication subcircuit implementation variants for CMOS ST circuits with a diphas operation discipline and researches the ways of their accelerating and simplifying. Coding discipline of the information signals is dual-rail in the combinational part and bi-phase (output of RS-trigger, [1]) in sequential part of the ST circuits.

The scientific novelty of the paper consists in researching feasible alternates of the multi-input hysteretic trigger that

* The study was done by a grant from the Russian Science Foundation (Project №. 19-11-00334)

speeds up and simplifies an indication subcircuit for any ST circuit, especially for multi-bit arithmetic digital units.

II. INDICATION SUBCIRCUIT IMPLEMENTATION BASIS

The classic principle of indicating circuits with dual-rail and bi-phase encoding of the information signals is as follows [2]:

- Generating a signal indicating spacer or work state of each data signal.
- Combining all internal indication signals into one output indication signal.

Signal indicating dual-rail information signal is generated using k-OR (k-NOR) cells for zero spacer or m-AND (m-NAND) cells for unit spacer, where k is a number of series-connected p-MOS transistors between supply bus and cell output, which is admissible in this process; m is a number of series-connected n-MOS transistors between ground bus and cell output, which is admissible in this process. Internal indication signals are combined by a pyramid circuit using the special indication cells.

Traditional indication cells are as follows: C-element (semi-static Muller's element) and hysteretic trigger (H-trigger, static Muller's element) [3]. Fig. 1 and 2 show 3-input CMOS circuits of the C-element and H-trigger respectively.

C-element uses the "weak" inverter (its transistors are outlined with dotted ovals in Fig. 1) to store its state between time moments, when all inputs of the C-element are the same. When C-element is forced to switch to an opposite state, a chain of sequentially connected "strong" p- or n-type transistors "draws" the potential of the internal net A supported by an opposite type transistor in "weak" inverter.

Advantages of the C-element are as follows:

- Small number of transistors: $2 \cdot (N + 2)$, where N is the number of inputs.
- Unit capacitance on each input.

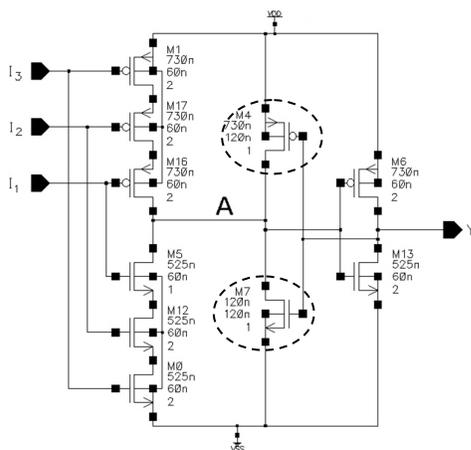


Fig. 1. 3-Input C-element

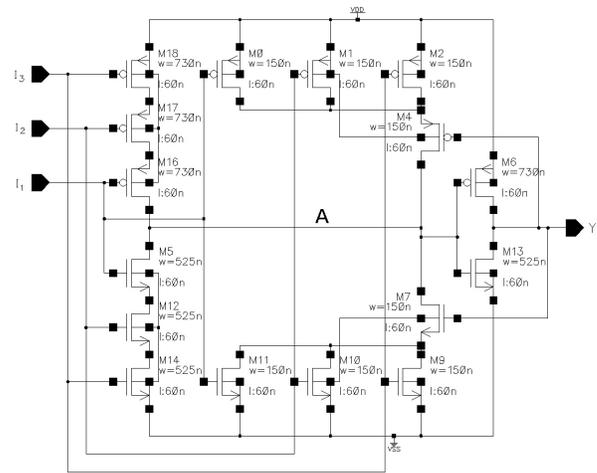


Fig. 2. 3-Input H-trigger

Disadvantages of the C-element are as follows:

- Availability of a short-circuit current (a few picoseconds in 65-nm CMOS process) flowing through the chain of series-connected "strong" transistors and transistor from "weak" inverter during switching C-element.
- Low noise immunity due to the fact that internal net A potential is supported by "weak" transistor; under the influence of strong enough interference it may change to a level switching "strong" output inverter and inverting state of the C-element.

In H-trigger, transistors providing holding trigger's state may have an arbitrary size because such an active transistor is disconnected from the power source or ground at a time when all inputs of the H-trigger take the same value not matching the stored state. Therefore, H-trigger switches in an opposite state without any process of "drawing" potential of the internal node A to new value.

Advantages of the H-trigger are an opposite to disadvantages of C-element:

- Lack of short-circuit current during switching.
- High noise immunity because the stored trigger's state is supported by "strong" transistors.

In addition, it has the better performance due to the lack of "drawing" the internal node A potential.

Disadvantages of the H-trigger are as follows:

- Increased number of transistors: $4 \cdot (N + 1)$, where N is the number of inputs.
- Input capacity is larger than in C-element.

Due to limitations on the number of series-connected transistors in CMOS circuit (no more than three p-MOS transistors and not more than four n-MOS transistors) indication subcircuit is based on 2-input and 3-input C-elements or H-triggers. Indication subcircuit combining M indication signals into single one can be implemented on

" $\lceil M \cdot (1 - 1/\log M) \rceil$ " 2-input C-elements and H-triggers or on " $\lceil \frac{M}{2} \cdot (1 - 1/\log_3 M) \rceil$ " 3-input their versions. It will have " $\lceil \log M \rceil$ " or " $\lceil \log_3 M \rceil$ " layers (cascades) of such cells respectively.

For example, the number of indication signals at first Wallace "tree" layer of a double precision multiplier compliant to IEEE754 standard [4] using dual-rail with unit spacer encoding equals to 1431. The indication subcircuit combining them will have 716 3-input and 2-input H-triggers located on the 7 layers of a pyramidal structure. One H-trigger has roughly 50-ps delay in 65-nm CMOS process in typical conditions. Thus a total delay of such indication subcircuit will be around 350 ps.

In applications that do not require the maximum reduction of dynamic current consumption, it is permissible to use multi-input H-triggers [3, 5] whose behavior is described by a Boolean function:

$$Y^+ = I_1 * I_2 * \dots * I_N + Y^*(I_1 + I_2 + \dots + I_N),$$

where I_1, I_2, \dots, I_N are the inputs of the N-input H-trigger. Fig. 3 demonstrates CMOS circuit of the N-input H-trigger.

The peculiarities of the multi-input H-triggers are as follows:

- Lack of connected in series transistors controlled by the inputs.
- "Weak" inverter (marked by dashed oval) is controlled by one of the trigger's input (I_N in Fig. 3) rather than by its output.

Multi-input H-trigger is also semi-static. Premature switching I_N input to a value corresponding to the next phase of work of the H-trigger causes a short-circuit current. And this current lasts until all inputs of the H-trigger will switch to the same value as I_N . Short-circuit current strength depends on the width of both the transistors in the "weak" inverter and opposing it serial-parallel transistor group in the input part of the trigger.

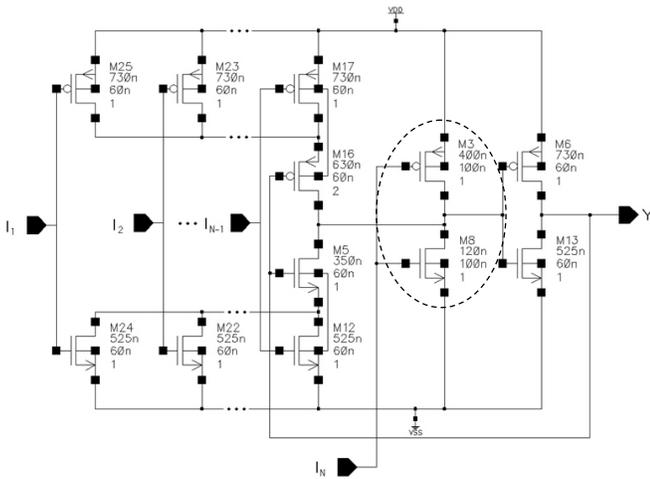


Fig. 3. N-input H-trigger

Circuit in Fig. 3 works correctly and without short-circuit current at any size of the transistors, if I_N input is changed the

most recent of all H-trigger inputs. To provide this logically, a designer can connect to this input an indication signal formed by longer cell chain than other indication signals. However, the behavior of the ST circuit should not depend on the delay of its elements. Consequently, one must to take into account that I_N may be either delayed in relation to the rest of the trigger inputs, or preceding switch of at least one of them, even if I_N propagates through longer cell chain. So there are specific requirements for the implementation and usage of the multi-input H-triggers.

III. MULTI-INPUT H-TRIGGER OPTIMIZATION

A necessary condition for the workability of the multi-input H-trigger is that "weak" inverter should not lead to a premature switching trigger at early changes of I_N , when at least one of the other inputs remained in a opposite state. In addition, the size of the transistors of the multi-input H-trigger should provide acceptable "performance to short-circuit current value" ratio in the worst case.

The necessary workability conditions for the multi-input H-trigger in typical 65-nm CMOS process are achieved with the following transistor size ratios:

$$\begin{cases} \frac{L_{p,weak}}{W_{p,weak}} \geq K_{p,GM} \cdot \left(\frac{L_{n,in}}{W_{n,in}} + \frac{L_{n,FB}}{W_{n,FB}} \right), \\ \frac{L_{n,weak}}{W_{n,weak}} \geq K_{n,GM} \cdot \left(\frac{L_{p,in}}{W_{p,in}} + \frac{L_{p,FB}}{W_{p,FB}} \right), \end{cases} \quad (1)$$

where $W_{p,weak}, W_{n,weak}, L_{p,weak}, L_{n,weak}$ are the width and length of p- and n-transistors in the "weak" inverter; $W_{p,in}, W_{n,in}, L_{p,in}, L_{n,in}$ are the width and length of p- and n-transistors driven by other H-trigger inputs; $W_{p,FB}, W_{n,FB}, L_{p,FB}, L_{n,FB}$ are the width and length of p- and n-transistors providing storing H-trigger's state at time intervals, when its inputs have the differential values; $K_{p,GM}, K_{n,GM}$ are coefficients depending on process-dependent parameters.

Simulation by means of Spectre program (Virtuoso, Cadence) has allowed for calculating coefficients $K_{p,GM}$ and $K_{n,GM}$ values for a standard 65-nm CMOS process. Taking into account the possible combinations of parameters of p- and n-transistors, they have been determined as $K_{p,GM} = 0.9$ and $K_{n,GM} = 6.4$. Size of the transistors in Fig. 3 matches the ratios (1).

Similarly, the transistor sizes in the C-element circuit are calculated to ensure proper operation of the C-element for all combinations of parameters of the p- and n-transistors and permissible ambient conditions. For example, for 3-input C-element:

$$\begin{cases} \frac{L_{p,weak}}{W_{p,weak}} \geq K_{p,c} \cdot \left(\frac{L_{n,in1}}{W_{n,in1}} + \frac{L_{n,in2}}{W_{n,in2}} + \frac{L_{n,in3}}{W_{n,in3}} \right), \\ \frac{L_{n,weak}}{W_{n,weak}} \geq K_{n,c} \cdot \left(\frac{L_{p,in1}}{W_{p,in1}} + \frac{L_{p,in2}}{W_{p,in2}} + \frac{L_{p,in3}}{W_{p,in3}} \right), \end{cases} \quad (2)$$

where $W_{p,in*}, W_{n,in*}, L_{p,in*}, L_{n,in*}$ are the width and length of the p- and n-transistors driven by the corresponding C-element input. For standard 65-nm CMOS process, taking into

account the possible combinations of the p- and n-transistor parameters, coefficients $K_{p,C}$ and $K_{n,C}$ have values $K_{p,C} = 0.7$ and $K_{n,C} = 8.1$.

Short-circuit current in the multi-input H-trigger depends on the order of switching its inputs. If I_N input driven the "weak" inverter switches last, the short-circuit current is absent, and vice versa, if it switches first among all inputs of the H-trigger, the short-circuit current is maximum.

Transistor sizes in the N-input H-trigger shown in Fig. 3 provide some balance between short-circuit current at worst condition and trigger's speed. At the same time, they ensure proper operation of the H-trigger at any switch order of its inputs. It is possible to improve performance by increasing the width of transistors in the "weak" inverter. But this will inevitably lead to an increase in the short-circuit current in a worst case.

Fig. 4 shows a family of diagrams presenting the short-circuit current I_S value in the circuit in Fig. 3 when the I_N input switches next-to-last, with various process-dependent parameters of transistors ("ff" – all transistors are "fast", "tt" – all transistors are typical, "ss" – all transistors are "slow"). For nominal supply voltage ($VDD = 1.0V$) the current I_S does not exceed $120 \mu A$ throughout the range of ambient temperature at any parameters ratio of the p- and n-transistors.

Short-circuit current in the C-element is comparable to the short-circuit current in the multi-input H-trigger. Its duration at fixed supply voltage and ambient temperature is determined only by the size ratio of "weak" and "strong" transistors and their parameters.

IV. COMPARISON OF INDICATION SUBCIRCUIT IMPLEMENTATIONS

Due to the nature of CMOS transistors operation at different temperatures and supply voltages, usage of the multi-input H-trigger is not always appropriate. Following are the results of simulating different variants of the indication subcircuit, combining specified number of the indication signals into a single signal.

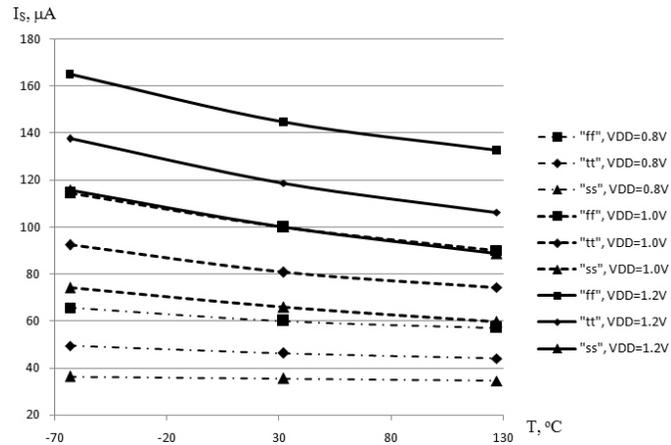


Fig. 4. Short-circuit current in 16-input H-trigger in the temperature and supply voltage VDD range

To compare performances of the different types of the indication elements, a ring oscillator was used. Fig. 5 shows its circuit. It consists of 10 identical segments (ISC) combining some indication signals and based on a "tree" of 2-input and 3-input indication cells (C-elements (C) or H-triggers (G)), or on a single multi-input H-trigger (GM), and one NAND2 cell enabling generation by signal $EN=1$. Contribution of the NAND2 to total generation period is insignificant.

Fig. 6 - 8 demonstrate the dependence of the generation period on supply voltage at various temperatures and ratios transistor parameters for three ring oscillators built of indication subcircuits combining 16 indication signals in different basis. Each type of curve corresponds to one corner of the transistor parameters: dotted line – "ss", dashed-dotted line – "tt", solid line – "ff". The results were obtained by means of program Spectre.

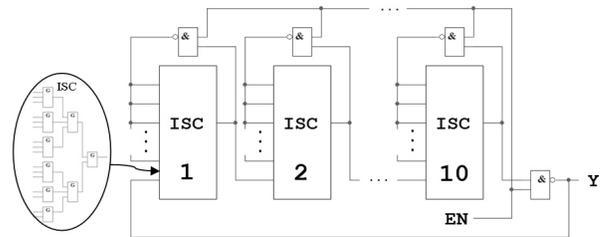


Fig. 5. Ring oscillator

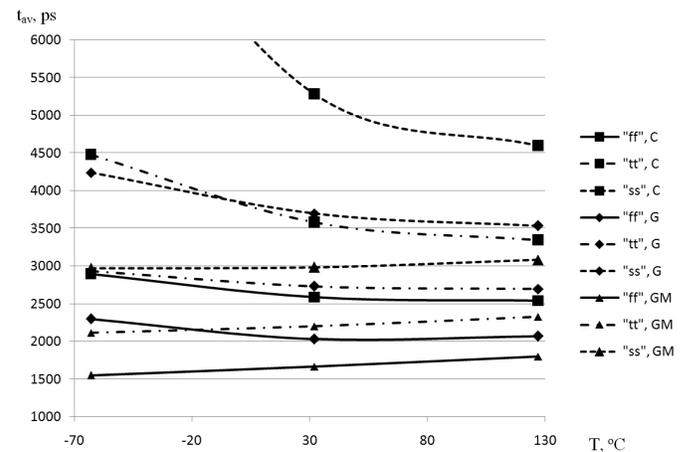


Fig. 6. Oscillation period of three ring oscillators for 0.8V supply voltage

Analysis of graphs in Fig. 6 – 8 shows the following:

- Indication subcircuit, combining 16 indication signals into a single output and implemented by one 16-input H-trigger, with supply voltages of 0.8V and 1.0V has the best performance compared with similar circuits on the C-elements and conventional H-triggers,
- At 1.2V supply voltage the advantage of multi-input H-trigger is restricted by the temperature range minus $63^{\circ}C$ through plus $50^{\circ}C$ for "ss" corner, and is preserved throughout full temperature range in all other corners, decreasing at increased ambient temperature,

- Performance of the indication subcircuit on base of C-element turned out to be worse among others variants under all conditions.

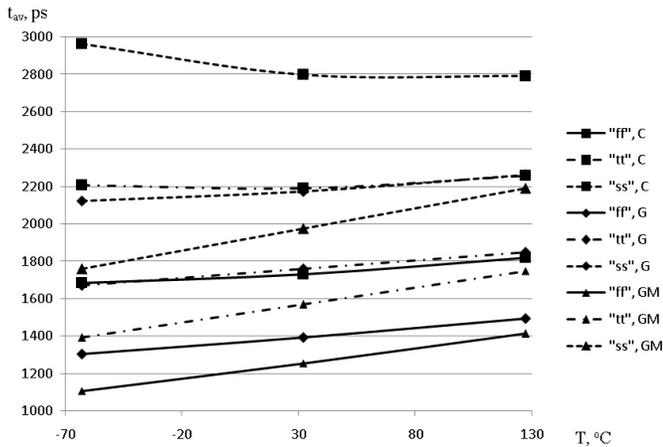


Fig. 7. Oscillation period of three ring oscillators for 1.0V supply voltage

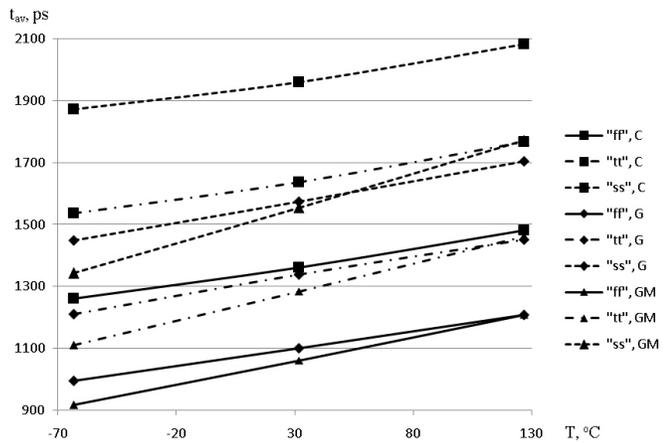


Fig. 8. Oscillation period of three ring oscillators for 1.2V supply voltage

Fig. 9 and 10 present the simulation results for indication subcircuits on base of H-trigger and multi-input H-trigger combining 9 and 27 indication signals at 1.0V supply voltage and in a range of temperatures and parameters of transistors. They show that multi-input H-trigger efficiency falls with decreasing number of combined indication signals. Indication subcircuit combining 9 signals and implemented on base of conventional 3-input H-triggers has better performance in the positive temperature range than 9-input H-trigger.

On the contrary, the 27-input H-trigger shows better performance throughout full temperature range than subcircuit on 3-input H-triggers. Its advantage almost linearly increases from (1...3)% at $T = 127^\circ\text{C}$ up to (17...24)% at a $T = -63^\circ\text{C}$ depending on transistor parameters.

Multi-input H-trigger shows similar advantage also at lower supply voltages. For example, at 0.8V supply voltage, its performance is better than performance of the subcircuit on base of conventional H-triggers by (13...14)% at $T = 127^\circ\text{C}$ and by (36...62)% at $T = -63^\circ\text{C}$ depending on transistor parameters. Increasing supply voltage reduces this advantage.

Moreover, in corner (1.2V supply voltage and $T = 127^\circ\text{C}$) this advantage disappears.

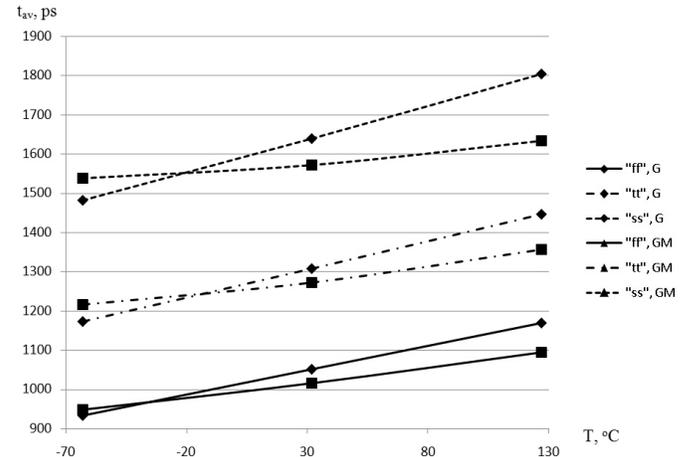


Fig. 9. Oscillation period of 9-input indication subcircuits for 1.0V supply voltage

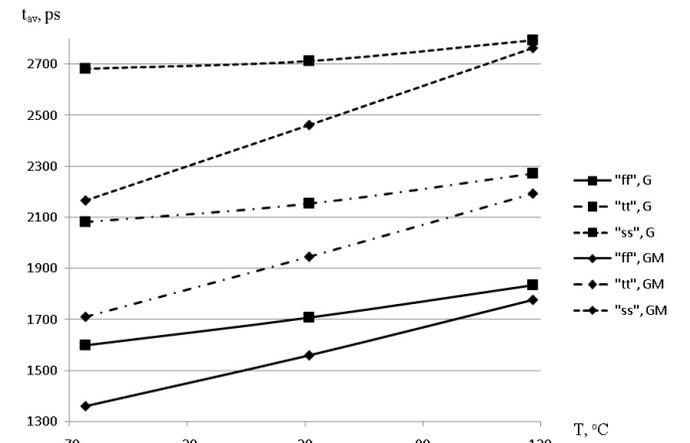


Fig. 10. Oscillation period of 27-input indication subcircuits for 1.0V supply voltage

However, multi-input H-trigger has additional advantages in comparison with traditional indication subcircuits on base of conventional H-triggers:

- Lower hardware costs (number of transistors N-input H-trigger equals to $2 \cdot (N + 2)$, that is identical to the formula for C-element),
- Simpler layout.

At the same time, the effectiveness of the multi-input H-triggers depends on the structure of the circuit generated indication signals to be combined. The maximum efficiency can be achieved in multiple circuits with almost simultaneous generation of bitwise indication signals: registers, parallel adders, parallel multipliers, etc. For example, the most balanced splitting multiplier 53×53 on two pipeline stages leads to appearing 598-bit intermediate register. Implementation of an indication subcircuit for this register on base of 30-input and 20-input H-triggers instead of 3-input H-triggers reduces its complexity by factor of 3.6, and

accelerates it by 14% at 1.0V supply voltage and 25°C ambient temperature.

Note that simulation results were obtained for the case the input of the multi-input H-trigger driving "weak" inverter changes later than other inputs in each H-trigger. This ensures the absence of any short-circuit current, but slightly slows down its work. Otherwise, multi-input H-trigger shall demonstrate higher performance, and short-circuit current will appear which value will correspond to the data shown in Fig. 4. This current will leak until all H-trigger's inputs switch to the same state.

Thus, the usage of multi-input H-triggers is appropriate for combining large number of the indication signals generated by bits of the parallel computing units and registers.

V. CONCLUSIONS

Indication subcircuit is a "bottleneck" of the multi-bit digital ST circuits. The need to detect the completion of the transitional processes in all elements of the ST circuit and to organize request-acknowledge interaction between ST blocks forces a developer to spend additional hardware and to slow down the circuit by forming indication subcircuit. It combines all internal indication signals into a single signal used as a control signal for preceding and subsequent blocks in the path of digital data processing.

Usage of the multi-input H-triggers in indication subcircuit of the multi-bit computing units and registers allows both for reducing hardware costs of the indication subcircuit implementation by several times, and for decreasing its delay by one and a half times, thereby increasing the performance of the entire ST circuit.

Under the typical values of the supply voltage ($V_{DD} = 1.0V$), ambient temperature ($T = 25^{\circ}C$) and model parameters of CMOS transistors ("tt" corner), 27-input H-trigger is faster (by 1.11 times) and less complex (by 3.6 times) in comparison

with similar indication subcircuit on base of 3-input H-triggers. Low voltage and low ambient temperature maximize a performance of the indication subcircuit using multi-input H-triggers.

C-element, which transistors are resized to provide the same short-circuit current during switching C-element as multi-input H-trigger has, demonstrates the worst performance in comparison with both conventional H-trigger and multi-input H-trigger.

Varying the size of transistors in the multi-input H-trigger circuit allows for shifting balance between its short-circuit current and performance in any direction/ One can accelerate H-trigger at the expense of increasing allowable short-circuit current or reduce possible short-circuit current, due to deterioration in its performance.

REFERENCES

- [1] Y.A. Stepchenkov, A.N. Denisov, Y.G. Diachenko, F.I. Grinfeld, O.P. Filimonenko, N.V. Morozov, et al. "Functional cell library for designing self-timed semi-custom chips on gate arrays 5503/5507". Moscow: Tekhnosfera. 2017. 367 p. — ISBN 978-5-94836-332-5. URL: <http://www.technosfera.ru/lib/book/497p>.
- [2] M. Kishinevsky, A. Kondratyev, A. Taubin, and V. Varshavsky. *Concurrent hardware: the theory and practice of self-timed design*, New York: J.Wiley & Sons, 1994, 368 p.
- [3] V.B. Marakhovskiy, "Theory of the logic design. Course slides," <http://elib.spbstu.ru/dl/1945.pdf/download/1945.pdf>. In Russian (last accepted date 17.05.2019).
- [4] IEEE Computer Society. 2008. IEEE Standard for Floating-Point Arithmetic IEEE Std 754-2008. doi:10.1109/IEEESTD.2008.4610935.
- [5] Y.A. Stepchenkov, Y.G. Diachenko, A.N. Denisov, and Y.P. Fomin. H-trigger. Patent № 2371842. Registered 27.10.09. Publ. in *Invention Bulletin*, 2009, № 30, 13p.

Main Solutions of Structural Health Monitoring in Managing the Technical Condition of Transport Objects

Andrei Belyi
Bridges Department
Emperor Alexander I St.
Petersburg State Transport
University
Saint-Petersburg, Russia
andbelyi@mail.ru

Dmitrii Shestovitskii
Bridges Department
Emperor Alexander I St.
Petersburg State Transport
University
Saint-Petersburg, Russia
shestovitsky@mail.ru

Eduard Karapetov
Bridges Department
Emperor Alexander I St.
Petersburg State Transport
University
Saint-Petersburg, Russia
eskar@yandex.ru

Dmitrii Sedykh
Department of Automation
and Remote Control on
Railways
Emperor Alexander I St.
Petersburg State Transport
University
Saint-Petersburg, Russia
sedyhdmitriy@gmail.com

Vladimir Linkov
Department of Automation,
Remote Control and
Communication on
Railway Transport
Russian University of
Transport (MIIT)
Moscow, Russia
linkov2@yandex.ru

ABSTRACT. Basis for operability of systems of tool monitoring are physical laws and algorithms of automation. The main used subsystems, such as control of the intense deformed state, vibration diagnostics, control of angles of rotation and inclinations of elements, are given in the text of article with the description of an essence of measurements. Standard schemes of constructions and their elements with the monitors located on them are also provided.

Keywords—structural health monitoring, monitoring subsystems, vibration diagnostics, technical condition

I. INTRODUCTION

Modern bridge constructions represent the difficult technical objects including simultaneously unique architectural decisions, and the advanced automation equipment of steering and control of state with further changes forecasting possibility. Development of scientific and technical progress allows in the course of creation of constructions and designs in the transport industry (bridges, flyovers, outcomes of roads in the different planes, road carpet, the railroads, etc.), to provide use of the integrated and external means of technical diagnosing and also expeditious periodic and continuous monitoring capable to quickly transfer data on technical condition of subject to diagnosing with the indication of the predicted terms of no-failure operation [1 – 4]. It considerably simplifies and reduces the price of difficult technical objects operation and also allows to increase the level of safety of their use.

Nowadays monitoring systems are the most adequate and precise instrument of civil objects elements diagnostic during their building and especially maintenance period of life cycle [5-7]. This tendency goes in parallel with new up-to-date approaches with global informative modeling and supervising of transport objects. Different examples of structural health monitoring realisation are transport objects:

bridges [8, 9], tunnels [10], embankments, locks and other constructions, as well as separate elements of transport infrastructure [11-15]. But the main emphasize among monitoring systems one can find in bridges realizations. Maybe it's the consequence of their "nature": frequently bridges constructions are located over big rivers, bays, artificial barriers.

In the seventies the 20th century monitoring of constructions came down to automatic instrumental data acquisition from keyzones of constructions. The situation slowly went from the systems loaded only by static parameters to up-to-date systems of monitoring (SHMS – Structural Health Monitoring System) on a "turn-key basis", with lots of test controllers and the built-in estimated analytical system. Such systems effectively developed in the EU countries, so what implementation is promoted by the legal framework allowing the infrastructure facility stakeholder to obtain decreasing coefficient in the cost of an obligatory insurance. In case of project cost many billions of euros even the tenth shares of percent of economy on an insurance premium are very considerable sums.

Besides the financial incentives the situation in the field of implementation of periodic and continuous monitoring means sharply changed as a result of breakthrough in development of information technologies in the last thirty years. High-precision gauges, modern blocks, digital converters, an optical wireless network, global positioning systems and other technical achievements laid a way to bigger accuracy, speed and profitability of data acquisition. For purposes of structural analysis the latest software is used that increases productivity of processing of large volume of data. Recently the significant contribution was made to increase in reliability of the equipment and mechanisms and environmental safety of constructions that affects also of their operation efficiency and service.

With development of the market of these services of SHMS it was selected as a separate class (services) necessary to application in the construction industry. In the market various technical solutions, both from domestic, and from foreign vendors appeared.

The described picture and situation is not local problem only. As one can see, a great deal of investigations go on all over the world [1-16]. In the present article authors give the information of St. Petersburg monitoring objects. With some features the city is a museum и bridges, thus the illustrating example is pretty evident.

That is why precisely city transport constructions are the objectives of the article. All the described data is obtained as a result of long period analysis, carried out by authors.

Modern systems of instrumental monitoring (structural health monitoring systems) have different physical principles. However and most commonly, all of them can be joined by measured parameters. This, in turn, will allow achieving an integral effect when the values are obtained from subsystems having a different basis.

As monitored parameters, the values obtained by direct measurements or indirectly, based on the results of direct measurements of other quantities that are functionally related to the desired quantity, can be used.

The goal of present article is to illustrate the necessity of different monitoring subsystems usage. For this goal some tasks are:

- indicating the parameters;
- pointing of subsystems;
- physical principles;
- examples objects.

During the monitoring of building structures of transport facilities, it is necessary to determine the needed parameters of objects' various parts. The structure elements (supports and span structures) subjected to the greatest loads and the greatest state changes in the course of construction and operation are subject to monitoring.

The main parameters to be monitored are:

- absolute and relative structures displacement;
- dynamic parameters affecting structures wearing process;
- stressed-deformed condition of structural elements;
- development of cracks (if any).

II. STRESSED-DEFORMED CONDITION

This condition is mainly fixed by strain gauge way with the help of special sensors (strain gauges). They fix the deformation at a certain point of the element, and then the stresses are determined using Hooke's Law. Deformations fixed on a segment called S base, during operating in the elastic stage, are registered by small values. Strain gauges measure the absolute elongation (shortening) ΔS and determine the average relative deformation:

$$\varepsilon = \frac{\Delta S}{S} \quad (1)$$

For the average relative deformation more accurately reflect the true one, S base has to be as small as possible.

In a linear stressed condition, it is enough to measure ΔS to determine the voltage on the base located in the acting force direction. Based on the obtained ε value and the known modulus of E elasticity, the stress is calculated:

$$\sigma = \varepsilon E \quad (2)$$

In case of a plane stressed condition at the given zone, deformations are measured in two or three directions.

Gauges are located along the main stresses σ_1 and σ_2 or (if the main stresses directions are unknown) one gauge can be set arbitrarily, and the other two can be set at angles of 45° and 90° or 60° and 120° to it. In the first case (the directions of main stresses are known), σ_1 and σ_2 are defined as follows:

$$\left. \begin{aligned} \sigma_2 &= \frac{E}{1-\mu^2} (\varepsilon_2 + \mu\varepsilon_1), \\ \sigma_1 &= \frac{E}{1-\mu^2} (\varepsilon_1 + \mu\varepsilon_2); \end{aligned} \right\} \quad (3)$$

where,

μ is Poisson's ratio.

In the second case, the calculations are more sophisticated, but can be made. To avoid excessive data in present article we won't post these information.

In the last decade of "advanced" strain gages – tensoresistors which principle of action is based on use of dependence between deformation and electrical quantities are more often used: ohmic resistance (mainly), capacity, inductance, etc. Deformation in the sensor causes change of one of electrical quantities which is measured with high precision; determine the amount of deformations by change of electrical quantity.

Important characteristic of the tensoresistor – tenczoefeling coefficient η :

$$\eta = \frac{\Delta R:R}{\varepsilon} \quad (4)$$

where,

R – nominal resistance of the tensoresistor;

ΔR – change of resistance of the tensoresistor;

ε – the relative deformation determined by formula (1).

Tensoresistors actually measure relative lengthening ε , but not change of length of ΔS base, as at strain gages.

The coefficient of a "tenczoefeling" depends on properties of material of which the tensoresistor, and technologies of its production is manufactured. This coefficient can differ from party to party (and even from one sensor to another) therefore it is always given by the manufacturer in the accompanying documents. The less difference in coefficient between series of sensors, the it is more qualitative and more reliable than measurement.



Fig. 1. Typical tensoresistor

Strain gages (tensoresistors) are sensors of measurement of deformation (tension). Allow to determine changes of tension by changes of electric characteristics. Can be established at any stage of construction (operation). The beginning of work of SHM during which indications of strain gages are considered as zero is considered the initial point. Control points of installation of strain gages are defined by calculation of bearing structures.

III. VIBRATION MONITORING SUBSYSTEM

It provides structures' dynamic parameters in the form of sets of accelerations and frequency patterns of oscillations. The parameters integrally contain data on structures' stiffness and masses, as well as external influences.

Measurements outcomes during dynamic monitoring allow revealing hidden changes in the structures strength properties. Therefore, tasks of dynamic monitoring include the next:

- dominant frequencies of free oscillations determination;
- seismic activity impact on the structure's dynamic operation assessment;
- determination of the level of transport loads influence on dynamic characteristics;
- frequency analysis to assess and predict changes in the technical condition.

The necessity to solve the tasks posed within the framework of dynamic monitoring opens a wide field both for researching the structures themselves, along with estimation of hidden damage development, and in terms of methods, tools, and setting monitoring tasks. In this regard, it should be noted that parameters of natural oscillations, presented as the set of frequencies and corresponding vibration modes, are the main characteristics of any design.

The next formula is well known based on the structures dynamics

$$(C - \lambda E)\vec{v} = 0 \quad (5),$$

where,

$C = AM$;

A – compliance matrix of a system with n degrees of freedom;

M – lumped mass matrix;

E – unitary diagonal matrix;

λ – matrix C eigenvalue;

\vec{v} – matrix C eigenvector.

The accelerometer serves for measurement of response characteristics of construction. Directly accelerometers measure acceleration in installation points. Acceleration is recalculated in other response and direct current characteristics, in particular, of vibration, natural frequencies, movements. For application as a part of the equipment of monitoring of bridge constructions accelerometers with the lower frequency from 0 Hertz are required.

Accelerometers have no start state.

Accelerometers are installed on constructions of a construction for registration of the fluctuations arising in it under the influence of different types of loadings (temporary, wind, seismic).



Fig. 2. Typical accelerometer

IV. SUBSYSTEM OF ANGLES AND OFFSETS CONTROL

The behavior analysis of the rod elastic line deformation under the influence of effects performed by the authors can be illustrated by Fourier Series with usage of trigonometric polynomials. In general and in this case, it is enough to use from six to eight members, depending on design features of the investigated structure.

Inclinometers should be located at the intersection points of the harmonics of Fourier Series with the rod axis.

The deformable rod shape under the influence of external factors is determined by the function obtained by solving simultaneous trigonometric equations, wherein the data obtained using the installed inclinometers are substituted for the arguments.

The length of L rod elastic line will correspond to half of the spatial period of the first harmonic of Fourier Series. Consequently, the length of this period is $2L$, and the lengths of T_i periods of all the harmonics of the series are determined by the following formula

$$T_i = \frac{2L}{i}, i = 1, 2, \dots \quad (6)$$

In this case, trigonometric polynomial takes the following form

$$y(x) = y_0 + \sum_{i=1}^n \left(y_{si} \sin \frac{\pi i x}{L} + y_{ci} \cos \frac{\pi i x}{L} \right) \quad (7)$$

It is recommendable to install the sensors at the points where the figures appearing in (6) the function, sine and cosine, take zero values. The abscissas of such points on the

elastic line of the beam for any harmonics can be found by the following formula

$$x_{ik} = \frac{kL}{2i}, \quad i = 1, \dots, n; k = 0, \dots, 2i \quad (8)$$



Fig. 3. Typical inclinometer

As one can see, on the fig. 3 there is an inclinometer, made by precision technology, as a geodetic one, with oil liquid inside. They are more expensive neither MEMs inclinometers, but more precise and have no drift of zero.

V. INTEGRAL OPERATION OF SUBSYSTEMS AND TYPICAL CONTROL POINTS

Thus, in the monitoring system there are completely different physical parameters, which integrated processing collectively gives an opportunity to get a reliable picture of the facilities technical condition. At the present stage of automation and electronics development, all transmitted signals are a stream of information vectors, which are processed using current theories and practical recommendations [16-19].

Besides named subsystems, there are some others – additional ones. For instance, meteostations and GNSS can be pointed.

Meteorological station represent the sensor of control of speed, the direction of wind, humidity, ambient temperature, pressure and intensity of rainfall. Are an integral part of a system of definition of a condition of designs for a possibility of assessment of integrated characteristics.

Typical schemes of parameters' control points are quite different depending on the belonging of a transport infrastructure object to a particular class: a bridge, a tunnel, etc. In addition, structures differ significantly in statistical schemes, construction material and other parameters. However, we will show some typical schemes of equipment layout (Fig. 4-10)¹.

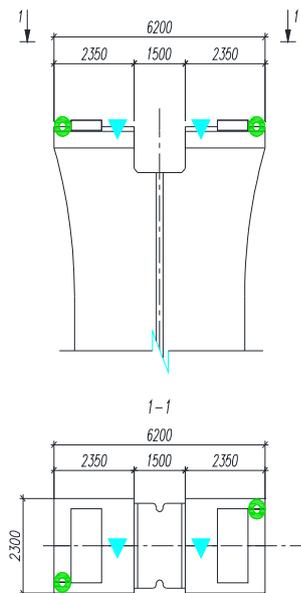


Fig. 4. Sensors layout on piers

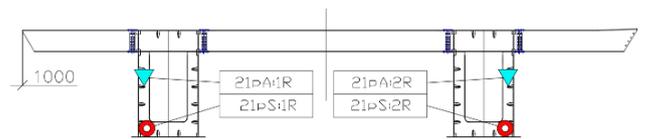


Fig. 5. Sensors layout on a metal span structure

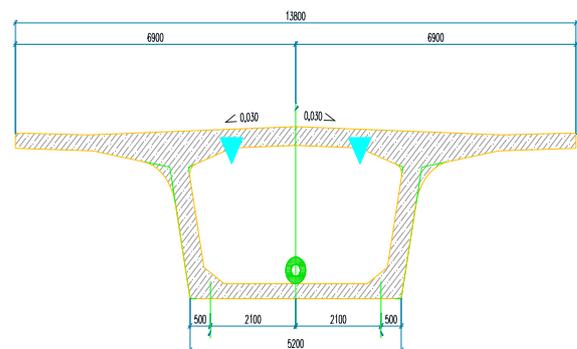


Fig. 6. Sensors layout on a reinforced concrete span

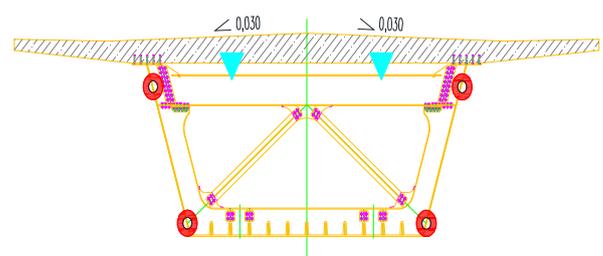


Fig. 7. Sensors layout on a steel-concrete superstructure

¹Note: On figures 4-9 red circle means strain gauge, green one – inclinometer, triangle is accelerometer.

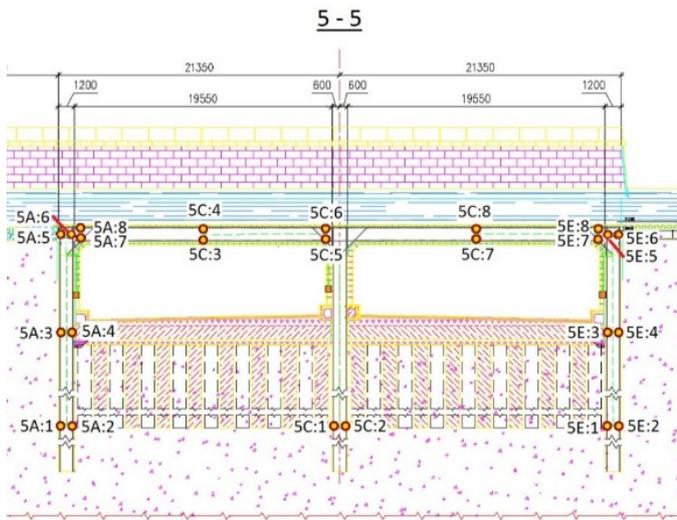


Fig. 8. Sensors layout on tunnel supports and walls

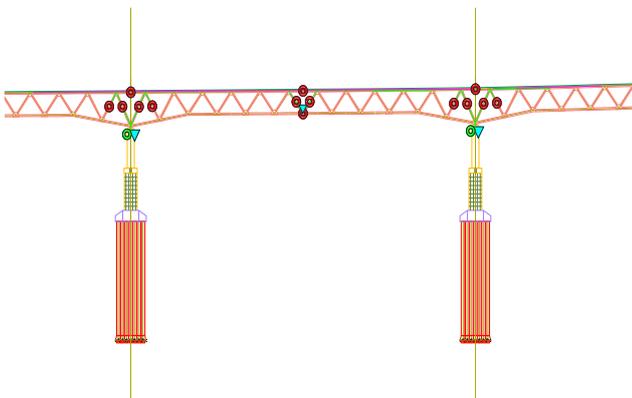


Fig. 9. Sensors layout on a lattice span

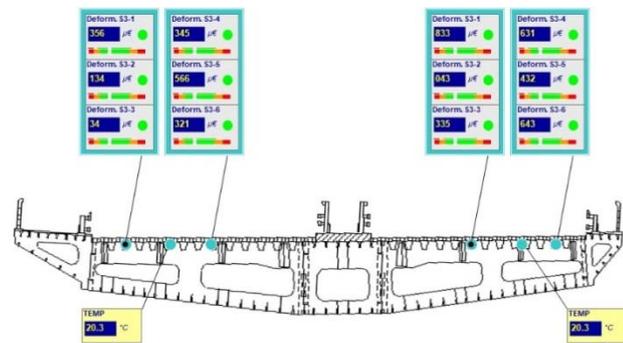


Fig. 10. Sensors on cross section in software complex

CONCLUSION

The applying of monitoring systems in all spheres of human activity is increasingly used in the processes of automating the management of the state of transport infrastructure facilities, such as bridges and tunnels.

Typical equipment schemes with monitoring devices were proposed for the main objects of the city, although it is required to develop full-fledged individual documentation for extra-curricular objects in each specific case.

The proposed solutions are based on the use of various subsystems of instrumental monitoring, which have different

physical bases. Using them in combination and integrally allows getting accurate and reliable results. The last sentence we would mark as a scientific novelty.

REFERENCES

- [1] Bonessio N., Lomiento G., Benzeni G. "Damage Identification Procedure for Seismically Isolated Bridges" // *Structural Control Health Monitoring*. – 2011. – Vol. 19. – Pp. 565–578, doi: 10.1002/stc.448.
- [2] Comisu C.-C., Taranu N., Boaca G., Scutaru M.-C. "Structural Health Monitoring System of Bridges" // *Procedia Engineering*. – 2017. – Vol. 199. – Pp. 2054-2059, doi: 10.1016/j.proeng.2017.09.472.
- [3] Alonso L., Barbarán J., J., Díaz M., Llopis L., Rubio B. "Middleware and Communication Technologies for Structural Health Monitoring of Critical Infrastructures: A Survey" // *Computer Standards & Interfaces*. – 2018. – Vol. 56. – Pp. 83-100, doi: 10.1016/j.csi.2017.09.007
- [4] Yi T.-H., Huang H.-B., Li H.-N. "Development of Sensor Validation Methodologies for Structural Health Monitoring: A Comprehensive Review" // *Measurement*. – 2017. – Vol. 109. – Pp. 200-2014, doi: 10.1016/j.measurement.2017.05.064.
- [5] Li J and Hao H. "Damage detection of shear connectors under moving loads with relative displacement measurements". *Mech Syst Signal Pr* 2015; 60–61: 124–150.
- [6] Lienhart, W., Ehrhart, M. "State of the art of geodetic bridge monitoring" *Structural Health Monitoring 2015: System Reliability for Verification and Implementation - Proceedings of the 10th International Workshop on Structural Health Monitoring, IWSHM 2015* DOI: 10.12783/SHM2015/58
- [7] Y. Yang, Q. S. Li, B. W. Yan. "Specifications and applications of the technical code for monitoring of building and bridge structures in China". *Advances in Mechanical Engineering*. – 2017. – Vol. 9 (1). p. 1–10. DOI: 10.1177/1687814016684272.
- [8] Mosbeh R. Kaloop and Jong Wan Hu. "Dynamic Performance Analysis of the Towers of a Long-Span Bridge Based on GPS" *Monitoring Technique/ Journal of Sensors* Volume 2016 (2016), Article ID 7494817, 14 pages <http://dx.doi.org/10.1155/2016/7494817>
- [9] H. Wenzel, *Health monitoring of bridges*, Chichester: John Wiley & Sons, 2009. – 621 p.
- [10] A.A. Belyi, E.S. Karapetov, Yu. S. Efimenko, "Structural health and geotechnical monitoring during transport objects construction and maintenance (Saint-Petersburg example)", *Procedia Engineering*. – 2017. – Vol. 189. p.145–151. PII:S1877-7058(17)32145-8 DOI:10.1016/j.proeng.2017.05.024
- [11] S. Sumitro, M.L. Wang., "Structural Health Monitoring System Applications in Japan," In: Ansari F. (eds) *Sensing Issues in Civil Structural Health Monitoring*. Springer, Dordrecht, 2005. – pp. 495-504. https://doi.org/10.1007/1-4020-3661-2_49.
- [12] J. E. Andersen, M. Fustinoni, *Structural health monitoring systems*, Italy: L&S S.r.l. Servizi Grafici, 2006. – 126 p.
- [13] W. Rucker, F. Hille, R. Rohrmann, *Guideline for structural health monitoring*. Final report, SAMCO, Berlin, 2006. – 63 p.
- [14] Efanov D., Osadchyy G., Sedykh D. "Development of Rail Roads Health Monitoring Technology Regarding Stressing of Contact-Wire Catenary System", In *Proceedings of 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, Chelyabinsk, Russia, 19-20 May, 2016, doi: 10.1109/ICIEAM.2016.7911431
- [15] Ana Paula Camargo Larocca, João Olympio De Araújo Neto, Jorge Luiz Alves Trabanco, Augusto César Barros Barbosa, André Luiz Barbosa Nunes Da Cunha, Ricardo Ernesto Schaal. *Uso de receptores GPS de 100 HZ na detecção de deflexões verticais milimétricas de pontes de concreto de pequeno porte*. *Bol. Ciênc. Geod.*, sec. Artigos, Curitiba, v. 21, no 2, p.290-307, abr-jun, 2015. <http://dx.doi.org/10.1590/S1982-21702015000200017> SCOPUS, 2015
- [16] Geoffrey R. Thomas, Akbar A. Khatibi. "Durability of structural health monitoring systems under impact loading". *Procedia Engineering* 188 (2017) 340–347 SCOPUS, 2017

- [17] Blagoveschenskaya E.A., Zuev D.V., Garbaruk V.V., Gerasimenko V.A., Sedykh D.V., Kunets D.S. "Application of convolutional neural networks for pattern recognition circuits of railway automatics. specifics of this application". In Proceedings of 2017 XX IEEE international conference on soft computing and measurements (SCM) 2017. P. 434-435.
- [18] Efanov D., Osadtchy G., Sedykh D. "Protocol of Diagnostic Information Transmission via Radio Channel Concerning Health Monitoring of Infrastructure of Russian Rail Roads". In Proceedings of 3ed International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), St. Petersburg, Russia, May 16-19, 2017.
- [19] Dmitry Sedykh, Denis Zuyev, Michael Gordon, Alexandr Skorokhodov, "Analysis of the Amplitude and Phase-Manipulated Signals of Automation Devices via Bluetooth Technology", In Proceedings of 2018 IEEE East-West Design and Test Symposium, EWDTS 2018.

A Technique for Semiconductor Devices Modeling Using Physical Templates

Alexandr M. Pilipenko
Department of Fundamentals of Radio
Engineering
Southern Federal University
Taganrog, Russia
ampilipenko@sfedu.ru

Vadim N. Biryukov
Department of Fundamentals of Radio
Engineering
Southern Federal University
Taganrog, Russia
vnbiryukov@yandex.ru

Alexander I. Serebryakov
JSC "Milandr"
Moscow, Zelenograd, Russia
sashaag@mail.ru

Abstract—A technique of template models creation for semiconductor devices (diodes and field-effect transistors) is developed in this paper. This technique is necessary for solving the problems of analog electronic circuits computer-aided design in robotic systems and space instrument engineering. The template model creation is carried out by the replacement of one or more parameters of the known physical model (template) by the relations of power series of control currents or voltages. The parametric identification of the template model is carried out by the method of least squares. The template model comprises all physical parameters of the initial model. The number of additional parameters in the template model is small (no more than four), which makes it possible to use standard minimum search algorithms for parametric identification. The obtained results show that the proposed template models provide the increase of the accuracy of modeling I-V characteristics of semiconductor devices more than twice in comparison with the known physical models.

Keywords—model, semiconductor devices, template, method of least squares, parametric identification

I. INTRODUCTION

One of the actual problems of robotics and space instrument engineering is the development of methods for simulation of integrated circuits (ICs) for processing the signals from sensors of various physical quantities [1]. Important elements of the aforementioned ICs are junction field-effect transistors (JFETs) which have a minimum level of the self-noise in a wide temperature range [2].

Currently, no universal models which ensure the acceptable accuracy for different temperatures and fabrication technologies exist for both JFETs and p-n junctions. For example, two compact physical models of p-n-junction are used in electronic circuit simulators: the first model is the model for the weak injection mode [3]; the second model takes into account the strong injection mode [4]. In some cases these compact models do not allow describing p-n junction characteristics with the acceptable accuracy [5] – [7].

The maximum accuracy of semiconductor devices basic characteristics modeling is achieved with use of table models [8]. A Table model is a set of numerical data represented in the

form of various arrays, as well as different methods of data interpolation. The main disadvantages of table models are as follows: the absence of physical parameters, a large volume of stored data, and the impossibility of describing the dependences of model parameters upon temperature.

High accuracy of semiconductor devices modeling can be achieved with use of template models which do not have the disadvantages of table models [9]. The essence of template modeling is that one or more parameters of the known physical model, which used as a template, are replaced by the function of control voltage or current.

The aim of this work is to develop universal technique for semiconductor devices template models creation with use of physical templates.

To attain the aforementioned aim the following problems are solved in the paper:

- development of accurate models of p-n junctions using templates;
- development of the parametric identification algorithm for semiconductor devices template models;
- proof of the efficiency of template models for approximation the I-V characteristics of p-n junctions made by various technologies.

II. DESCRIPTION OF KNOWN MODELS

The known equivalent circuit of a p-n junction (Fig. 1) is a series circuit which consists of the linear resistance R_S and the current source I_D with the intrinsic voltage control [10]. The equivalent circuit shown in Fig. 1 is used for description of p-n junction I-V characteristics in all SPICE- simulators.

Fig. 2 shows the measured I-V characteristics of two different p-n junctions [11]. The curve 1 was measured for the MURD315 diode (it corresponds to the weak injection mode). The curve 2 was measured for the FR102 diode (it corresponds to the strong injection mode).

A p-n junction current in the weak injection mode, when the concentration of minority charge carriers in the p- and n- regions is small, can be represented in the following form:

The reported study was supported by the Grant of the Russian Science Foundation according to the research project No. 16-19-00122-P

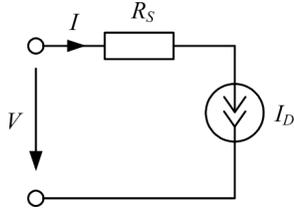


Fig. 1. Equivalent circuit of p-n junction.

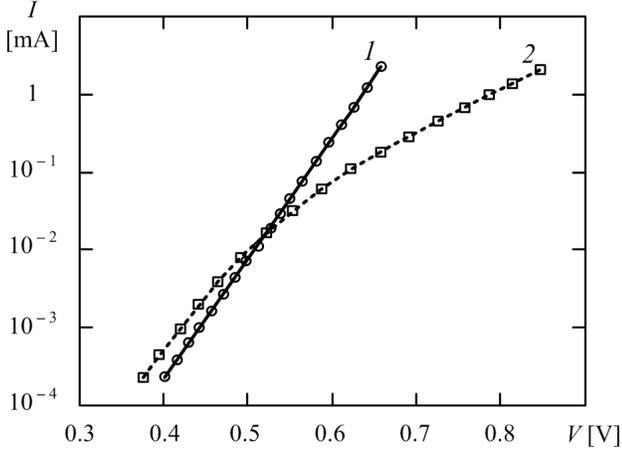


Fig. 2. Measured I-V characteristics of the p-n junctions with the weak injection (curve 1) and the strong injection (curve 2).

$$I = I_S \left[\exp\left(\frac{V - IR_S}{N\phi_T}\right) - 1 \right], \quad (1)$$

where V is the p-n junction voltage; I_S is the saturation current; N is the emission coefficient; $\phi_T = kT/q$ is the thermal voltage; $k \approx 1.38 \cdot 10^{-23}$ J/K is the Boltzmann constant; $q \approx 1.6 \cdot 10^{-19}$ C is the value of the elementary charge; T is the absolute temperature in Kelvin degrees.

The p-n junction current in the strong injection mode, which occurs under high concentration of minority charge carriers in the p- and n- regions, is described by the following expression

$$I = I_S \left[\exp\left(\frac{V - IR_S}{N\phi_T}\right) - 1 \right] \times \sqrt{\frac{I_{KF}}{I_{KF} + I_S \left[\exp\left(\frac{V - IR_S}{N\phi_T}\right) - 1 \right]}}, \quad (2)$$

where I_{KF} is the current that corresponds to the transition from the weak injection to the strong injection.

It should be noted that the model (1), which does not take into account the strong injection effects, is a special case of the model (2) where $I_{KF} \rightarrow \infty$.

The parameters of the compact models (1) and (2) are determined from the results of direct measurements, so these models can be used as the physical templates for the development of more accurate models of p-n junctions [9].

III. TECHNIQUE OF TEMPLATE MODELS CREATION

The known approach to template models creation is described in [12]. This approach consists in the expansion of the original physical model parameters into a power series of the control voltage or current. It was shown in [8] and [9] that the replacement of semiconductor devices models parameters by the relation of power series is more effective for template models creation than the expansion of parameters into a power series. The use of the relation of power series for modeling allows ensuring the monotonicity of the I-V characteristic, in contrast to the expansion into a power series, in which a violation of monotonicity is possible [13].

It is proposed in this paper to create a template model by replacing one or more parameters of the initial physical model by the relation of a power series:

$$P(x) = P_0 \frac{1 + b_1 x + b_2 x^2 + \dots + b_m x^m}{1 + a_1 x + a_2 x^2 + \dots + a_n x^n}, \quad (3)$$

where P_0 is the measured parameter of a physical model; x is a control voltage or current; a_1, a_2, \dots, a_n и b_1, b_2, \dots, b_n are the empirical coefficients; m is the order of the polynomial in the numerator; n is the order of the polynomial in the denominator.

As a rule, $n > m$, so the value of n determines the order of a template model. It is proposed in this paper to use template models of the first or the second order. In these cases the number of additional empirical coefficients is one or three, respectively, so the problem of parametric identification is complicated slightly.

As we know the resistance R_S depends nonlinearly on the current of a p-n junction [2], therefore it is proposed to replace this parameter by the relation of the power series:

for $n = 1$

$$R_S(I) = \frac{R_{S0}}{1 + a_1 I}; \quad (4)$$

for $n = 2$

$$R_S(I) = R_{S0} \frac{1 + b_1 I}{1 + a_1 I + a_2 I^2}. \quad (5)$$

It should be noted that the additional increase in the template model accuracy can be achieved by replacing the other parameters of the initial physical model by the expressions similar to (4) or (5).

IV. ALGORITHM OF PARAMETRIC IDENTIFICATION

The parameters of the p-n junction models are determined by the least-squares method from the minimum of the objective function [14]

$$S = \sum_{k=1}^M \left[\frac{I(V_k) - I_k}{I_k} \right]^2, \quad (6)$$

where M is the number of the experimental points; I_k and V_k are the measured values of the current and the voltage respectively; $I(V_k)$ are the current values calculated using the p-n junction model at $V = V_k$.

To solve the problem of the objective function minimum search we used Levenberg–Marquardt algorithm with the error control by variation of initial conditions.

To increase the speed of the objective function minimum search the parameters of the initial physical model were used as the initial conditions for parametric identification of the first order template model. The parameters of the first order template model, in their turn, were used as initial conditions for parametric identification of the second order template model. The initial values of the empirical coefficients were chosen to be zero.

The accuracy of modeling was estimated using different types of errors described below.

1. The relative error of the model at each point of the I-V characteristic:

$$\delta_k = \frac{I(V_k) - I_k}{I_k}.$$

2. The maximum relative error of the model:

$$\delta_{\max} = \max |\delta_k|.$$

3. The relative root-mean-square (RMS) error:

$$\sigma = \sqrt{\frac{S_{\min}}{M}},$$

where S_{\min} is the minimum value of the objective function.

To improve the accuracy of the parametric identification we recommend using the modified random search algorithm [15].

V. RESULTS OF MODELING

Table 1 shows the results of parametric identification of the p-n-junctions models with the weak and the strong injection (the measured I-V characteristics of the p-n-junctions are shown in Fig. 2).

The model (1) was chosen as the initial physical template for modeling the p-n junction with the weak injection, because in this case the model (1) provides approximately the same accuracy as more complex model (2) [11]. The template models (1) & (4) and (1) & (5) were obtained on the basis of the model (1) by replacing the R_S parameter by the expressions (4) and (5) respectively.

The model (2) was chosen as the initial physical template for modeling the p-n junction with the strong injection, because in this case the model (2) is an order of magnitude more accurate than the model (1) [11]. The template models (2) & (4) and (2) & (5) were obtained on the basis of the model (2) by replacing the R_S parameter by the expressions (4) and (5) respectively.

As we can see from Table 1 the use of the template models provides the decrease of the maximum and RMS errors of modeling approximately in 2 – 4 times in comparison with the initial physical model both for the p-n junction with the strong injection and the p-n-junction with the weak injection.

It should be noted that the second-order template model ($n = 2$) makes it possible to reduce the errors of modeling the p-n junction with the weak injection approximately twice in comparison with the first-order template model ($n = 1$). The second-order template model for the p-n junction with the strong injection is more accurate than the first order template model only in 1.2 times.

Fig. 3 illustrates the dependences of the relative errors of modeling the p-n junctions I-V characteristics upon the control

TABLE I. RESULTS OF PARAMETRIC IDENTIFICATION

Model	Model parameters							Model errors	
	I_S [pA]	N	R_S [Ohm]	I_{KF} [uA]	a_1 [mA ⁻¹]	a_2 [mA ⁻²]	b_1 [mA ⁻¹]	δ_{\max} [%]	σ [%]
<i>p-n junction with the weak injection</i>									
(1)	0.0158	1.001	0.621	–	–	–	–	1.66	0.71
(1) & (4)	0.0218	1.015	0.332	–	–0.036	–	–	0.69	0.41
(1) & (5)	0.0229	1.017	0.465	–	–0.037	–0.0026	–0.058	0.36	0.19
<i>p-n junction with the strong injection</i>									
(2)	1.425	1.222	23.55	17.4	–	–	–	6.68	2.93
(2) & (4)	1.043	1.195	55.17	16.8	0.66	–	–	2.47	1.19
(2) & (5)	1.133	1.202	25.12	16.6	4.90	4.04	11.4	2.00	1.01

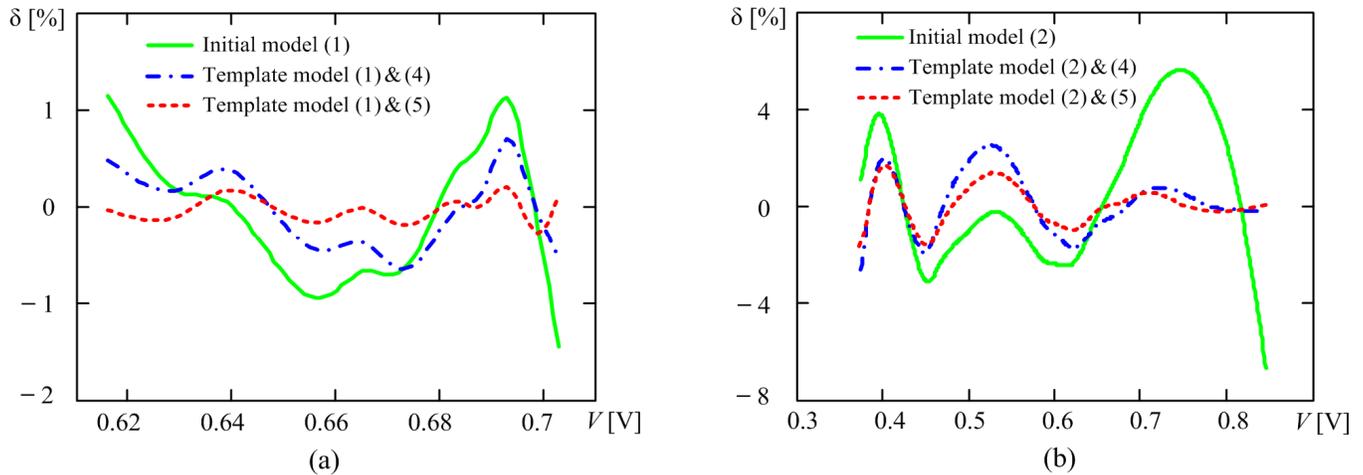


Fig. 3. Relative errors of modeling the p-n junctions with the weak injection (a) and the strong injection (b).

voltage. The results shown in Fig. 3 confirm the effectiveness of the template models use in the whole operating area of the I-V characteristic.

VI. CONCLUSIONS

A universal technique of semiconductor devices template models creation on the basis of the known physical models is developed in this paper. This technique is applicable for different types of p-n junctions, and as well as for JFETs and MOSFETs [8], [9]. Template models provide the increase of accuracy of I-V characteristics modeling in 2 – 4 times in comparison with the known physical models.

The proposed algorithm of parametric identification is realized on the basis of standard methods for the objective function optimization. This algorithm allows to obtain the error in determining the parameters of semiconductor devices models comparable to the error of I-V characteristics measurement.

The template model comprises all physical parameters of the initial model, so it can be used for semiconductor devices modeling in a wide temperature range [16], [17].

REFERENCES

- [1] O. V. Dvornikov, V. L. Dzialau, N. N. Prokopenko, K. O. Petrosiants, N. V. Kozhukhov, and V. A. Tchekhovski, "The accounting of the simultaneous exposure of the low temperatures and the penetrating radiation at the circuit simulation of the BiJFET analog interfaces of the sensors," 2017 International Siberian Conference on Control and Communications (SIBCON). Proceedings, 2017, doi:10.1109/SIBCON.2017.7998507.
- [2] S. S. Li, *Semiconductor Physical Electronics*, 2nd ed. Springer, 2006.
- [3] A. Vladimirescu, *The SPICE book*. John Wiley & Sons, 1994.
- [4] G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE*, 2nd ed. McGraw-Hill, 1993.
- [5] A. Baskys, M. Sapurov, and R. Zubavicius, "The New Equations of p-n Junction Carrier Injection Level," *Elektronika ir elektrotechnika*, vol. 19, no. 2, pp. 45–48, 2013, doi: 10.5755/j01.eee.19.2.3467.
- [6] A. Ortiz-Conde and F. J. García Sánchez, "Extraction of non-ideal junction model parameters from the explicit analytic solutions of its I-V characteristics," *Solid-State Electronics*, vol. 49, no. 3, pp. 465–472, March 2005, doi: 10.1016/j.sse.2004.12.001.
- [7] A. Ferhat-Hamida, Z. Ouennoughi, A. Hoffmann, and R. Weiss "Extraction of Schottky diode parameters including parallel conductance using a vertical optimization method," *Solid-State Electronics*, vol. 46, no.5, pp. 615–619, May 2002, doi: 10.1016/S0038-1101(01)00337-9.
- [8] V. N. Biryukov and A. M. Pilipenko, "Measurement-Based MOSFET Model for Helium Temperatures," *Proceedings of 2015 IEEE East-West Design and Test Symposium (EWDTS)*, 2015, pp. 241–244, doi: 10.1109/EWDTS.2015.7493109.
- [9] V. N. Biryukov, "Template modeling of a p-channel MOSFET," *Zhurnal Radioelektroniki – Journal of Radio Electronics*, no. 2, February 2019. doi: 10.30898/1684-1719.2019.2.11.
- [10] M. Khalis, R. Masrour, Y. Mir, and M. Zazoui "Two methods for extracting the parameters of a nonideal diode," *International Journal of Physical Sciences* vol. 10(8), pp. 270–275, April 2015, doi: 10.5897/IJPS2015.4260.
- [11] V. N. Biryukov and A.M. Pilipenko, "Diagnostics of the Nonlinear Static Models of a Diode," *Journal of Communications Technology and Electronics*, vol. 54, no. 5, pp. 577–582, May 2009, doi: 10.1134/S1064226909050118.
- [12] S. Van den Bosch and L. Martens "Approximation of State Functions in Measurement-Based Transistor Model," *IEEE Transactions on Microwave Theory and Techniques*, vol. 47, no. 1, pp. 14–17, January 1999, doi: 10.1109/22.740069.
- [13] C. C. McAndrew, "Practical modeling for circuit simulation," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 3, pp. 439–448, March 1998, doi: 10.1109/4.661209.
- [14] A. Ortiz-Conde, Y. Ma, J. Thomsonc, E. Santos, J. J. Lioub, F. J. García Sánchez, M. Lei, J. Finol, and P. Layman "Direct extraction of semiconductor device parameters using lateral optimization method," *Solid-State Electronics*, vol. 43, vo. 4, pp. 845–848, April 1999, doi: 10.1016/S0038-1101(99)00044-1.
- [15] A. M. Pilipenko and V. N. Biryukov, "Efficiency improvement of the random search algorithm for parametric identification of electronic components models," 2016 International Siberian Conference on Control and Communications (SIBCON). Proceedings, 2016, doi: 10.1109/SIBCON.2016.7491703.
- [16] H. A. Mantooth, R. G. Perry, and J. L. Duliere, "A Unified Diode Model for Circuit Simulation," *IEEE Transactions on Power Electronics*, Vol. 12, No. 5, Sep 1997, pp. 816–823, doi: 10.1109/63.622999.
- [17] A. M. Pilipenko and V. N. Biryukov, "Modeling of MOSFETs Parameters and Volt-Ampere Characteristics in a Wide Temperature Range for Low Noise Amplifiers Design," *Proceedings of IEEE East-West Design & Test Symposium (EWDTS)*, 2014, pp. 156–159, doi: 10.1109/EWDTS.2014.7027065.

Statistical Analysis of Discriminators under the Influence of Additive Correlated non-Gaussian Noise Described by Markov Processes

Vladimir Mikhaylovich Artyushenko
Information technology and management systems department
Technological University
Korolev city, Russian Federation
artuschenko@mail.ru

Vladimir Ivanovich Volovach
Informational and electronic service department
Volga Region State University of Service
Togliatty city, Russian Federation
volovach.vi@mail.ru

Abstract—To improve the accuracy of the evaluation of information parameters of the signal allows the function of the probability density of noise affecting the processed signal and errors of mismatch. The paper presents a statistical analysis of discriminators of nonlinear blocks of motion parameters meters under the influence of additive mixture of useful signal and correlated non-Gaussian noise. The synthesis of a discriminator, exposed to additive noise described by singly connected and multiply connected Markov processes and noise with independent values, is performed. Structural diagrams of the discriminators are shown, and the mathematical relationships describing the main characteristics of the discriminators, including the generalized discriminator, are given. The efficiency of the discriminator in the conditions of shifted hypotheses is analyzed. The main characteristics of the discriminator exposed to correlated additive noise with the finite mismatch error are obtained. It is noted that in the case of exposure to non-Gaussian noise, its reduction occurs due to decorrelation and nonlinear amplitude suppression, while, as a rule, it must be consistent with the spectral power density of the influencing noise.

Keywords—probability density function, non-Gaussian additive correlated noise, singly connected Markov process, generalized discriminator, mismatch error.

I. INTRODUCTION

Questions of synthesis and statistical analysis of discriminators of nonlinear blocks of meters of information parameters of signals are considered in a significant number of works [1-4, etc.], which usually took into account only the form of the signal. However, it is important, as will be shown below, to take into account the influence of the probability density function (PDF) of the influencing interference $W_n(n)$ and the error of mismatch $W(\epsilon)$.

Let us use the results of [5-10, etc.], which describe the effect on the signal of non-Gaussian additive noise with independent values.

Let the samples $\{y_h\}$ formed by the additive mixture of the processed signal $s(\lambda_h, t_h)$ (where λ_h is the information parameter of the signal) and the non-Gaussian noise with independent values enter the discriminator:

$$y_h = s(\lambda_h, t_h) + n_h, \quad h = \overline{1, H}. \quad (1)$$

The estimation equations for the case of high demodulation accuracy satisfying the maximum a posteriori probability criterion have the following form [6]:

$$\hat{\lambda}_h = \hat{\lambda}_{e,h} + \hat{\sigma}_{e,h}^2 \left[\hat{\lambda}_{e,h}, y_h - s(\hat{\lambda}_{e,h}) \right] B'_{\lambda,h};$$

$$\hat{\sigma}_{e,h}^2 \left[\hat{\lambda}_{e,h}, y_h - s(\hat{\lambda}_{e,h}) \right] = \left\{ \frac{\partial^2}{\partial \lambda_h^2} \ln W(\lambda_h | \lambda^{h-1}) - B''_{\lambda,h} \right\}^{-1} \Big|_{\lambda_h = \hat{\lambda}_{e,h}}$$

with $\hat{\sigma}_{e,h}^2 \left[\hat{\lambda}_{e,h}, y_h - s(\hat{\lambda}_{e,h}) \right]$ is the variance of the posteriority distribution of the estimation; $\hat{\lambda}_{e,h}$ is the extrapolated value λ for the h -th step, $B''_{\lambda,h}$ is the derivative $B'_{\lambda,h}$ with respect to the information parameter λ .

The output effect of the discriminator will be described by the expression:

$$B'_{\lambda,h} = s'_\lambda(\hat{\lambda}_{e,h}) z(n_{\Sigma,h}),$$

in which $s'_\lambda(\hat{\lambda}_{e,h})$ is a derivative of the signal relative to the measured parameter λ .

The characteristic of the nonlinear block (NLB), depending on the type of PDF noise $W_n(n)$ and the error of mismatch $W(\epsilon)$, is determined by the expression:

$$\begin{aligned} z(n_{\Sigma,h}) &= -\frac{d}{dn_{\Sigma,h}} \ln W_n \left[y_h - s(\hat{\lambda}_{e,h}, t_h) \right] = \\ &= -\frac{d}{dn_{\Sigma,h}} \ln W_n \left[s(\lambda_{e,h}, t_h) + n_{n,h} s(\hat{\lambda}_{e,h}, t_h) \right] = \\ &= -\frac{d}{dn_{\Sigma,h}} \ln W_n(n_{e,h} + n_h) = -\frac{d}{dn_{\Sigma,h}} \ln W_n(n_{\Sigma,h}), \end{aligned}$$

in which $n_{e,h} = s(\lambda_h, t_h) - s(\hat{\lambda}_{e,h}, t_h)$ there is a difference between the received signal and the reference signal, this difference can be both deterministic and random.

In the expression for logarithm of likelihood function (LLF) $\ln W_n \left[y_h - s(\hat{\lambda}_{e,h}, t_h) \right] = B_n(n_h)$ the difference contains information about the parameters of the useful signal, which the discriminator allocates.

The purpose of the work is that in real conditions, when the signal is under the influence of non-Gaussian noise, the condition $z(n_{\Sigma,h}) \approx z(n_{n,h})$ is not met and it is of practical interest to analyze the characteristics of the discriminator with unequal zero error misalignment, that is, in the presence of constant and random detuning between the input $s(\lambda_h, t_h)$ and the reference value $s(\hat{\lambda}_{e,h}, t_h)$ of the signal with extrapolated evaluation of the information parameter.

Note also that the output effect of the discriminator can be defined [6] as the sum of two random components: fluctuation $\alpha(t_h) = s'_\lambda(\hat{\lambda}_{e,h})z'_n(n_{n,h})$ and discrimination $\beta(\varepsilon) = \varepsilon_h \left[s'_\lambda(\hat{\lambda}_{e,h}) \right]^2 z'_n(n_{n,h})$ (DC) characteristics of the discriminator.

II. SYNTHESIS OF THE DISCRIMINATOR UNDER THE INFLUENCE OF ADDITIVE NOISE DESCRIBED BY SINGLE-CONNECTED MARKOV PROCESSES

Let it be assumed that non-Gaussian additive noise influencing the signal is a singly-connected continuous stationary Markov sequence with a given transition PDF $W_n(n_{n,h}|n_{n,h-1})$. In this case, the LLF will be described by the ratio

$$B_n(n_{n,h}|n_{n,h-1}) = \ln W_n \left[\left(y_h - s(\hat{\lambda}_h) \right) \middle| \left(y_{h-1} - s(\hat{\lambda}_{h-1}) \right) \right].$$

The assumptions about the characteristics of the processed signal $s(\hat{\lambda}_h, t_h)$ are given above.

Then the characteristic of the discriminator will be described by the expression

$$\begin{aligned} B'_\lambda(n_{n,h}|n_{n,h-1}) &= s'_{\lambda,h}(\hat{\lambda}_h, t_h) z_{n,h}(n_{n,h}) + \\ &+ s'_{\lambda,h-1}(\hat{\lambda}_{h-1}, t_{h-1}) z_{n,h-1}(n_{n,h-1}) = \\ &= s'_{\lambda,h} z'_{n,h} + s'_{\lambda,h-1} z'_{n,h-1}, \end{aligned} \quad (2)$$

where

$$\begin{aligned} z_{n,h} &= -\frac{\partial}{\partial n_{n,h}} \ln W_n(n_{n,h}|n_{n,h-1}); \\ z_{n,h-1} &= -\frac{\partial}{\partial n_{n,h-1}} \ln W_n(n_{n,h}|n_{n,h-1}). \end{aligned}$$

In contrast to the case of additive non-Gaussian noise with independent values, the $z_{n,h}$ and $z_{n,h-1}$ functions describe the characteristics of inertial nonlinear block (INLB).

Taking mismatch error $n_{\varepsilon,h}$, into account, in view of the above, $z_{n,h}$ and $z_{n,h-1}$ are assumed to be

$$\begin{aligned} z_{n,h} &= -\frac{\partial}{\partial n_{\Sigma,h}} \ln W_n(n_{n,h} + n_{\Sigma,h} | n_{n,h-1} + n_{\Sigma,h-1}); \\ z_{n,h-1} &= -\frac{\partial}{\partial n_{\Sigma,h-1}} \ln W_n(n_{n,h} + n_{\Sigma,h} | n_{n,h-1} + n_{\Sigma,h-1}). \end{aligned}$$

In this case, the diagram of the discriminator corresponding to algorithm (2) is shown in Fig. 1.

Studies have shown that when signals affected by correlated additive noise are processed, the input mixture is decorrelated in the discriminator. The decorrelator of noise in INLB provides suppression of correlated additive noise that is called frequency suppression. The effectiveness of optimal or real frequency suppression is usually assessed by the signal-to-noise ratio at the output and input of the decorrelator.

If the useful signal is affected by Gaussian correlated noise, then the block diagram of the discriminator will

contain two linear channels, in each of them there occurs the process of decorrelation.

If noise is non-Gaussian, it is suppressed by both decorrelation and nonlinear amplitude suppression. In general, linear decorrelator of noise should be coordinated with spectral density of the power of noise.

The main characteristics of the discriminator are considered and analyzed.

Using the approach for uncorrelated additive noise described earlier in [11, 12], we write down the output effect of the discriminator with respect to the mismatch error as

$$\varepsilon = \lambda_h - \hat{\lambda}_h.$$

Using the expansions:

$$z_{\varepsilon,h} \approx z_{n,h} + \varepsilon s'_\lambda(\hat{\lambda}_h) z'_{n,h,h} + \varepsilon s'_\lambda(\hat{\lambda}_{h-1}) z'_{n,h,h-1};$$

$$z_{\varepsilon,h-1} \approx z_{n,h-1} + \varepsilon_h s'_\lambda(\hat{\lambda}_h) z'_{n,h,h} + \varepsilon_{h-1} s'_\lambda(\hat{\lambda}_{h-1}) z'_{n,h-1,h-1}.$$

where $z_{n,i,j} = \frac{d}{d\hat{\lambda}_i} z_{n,i}$; $i, j = h, h-1$,

we write:

$$\begin{aligned} B'_\lambda(n_{n,h}|n_{n,h-1}) &\approx s'_\lambda(\hat{\lambda}_h) z_{n,h}(n_{n,h}) + s'_\lambda(\hat{\lambda}_{h-1}) z_{n,h-1}(n_{n,h-1}) + \\ &+ \varepsilon \left\{ \left(s'_\lambda(\hat{\lambda}_h) \right)^2 z'_{n,h,h} + s'_\lambda(\hat{\lambda}_h) s'_\lambda(\hat{\lambda}_{h-1}) z'_{n,h,h-1} + \right. \\ &+ \left. \left(s'_\lambda(\hat{\lambda}_{h-1}) \right)^2 z'_{n,h-1,h-1} + s'_\lambda(\hat{\lambda}_h) s'_\lambda(\hat{\lambda}_{h-1}) z'_{n,h-1,h} \right\} = \\ &= \alpha_h(n_h) + \beta_h(\varepsilon). \end{aligned}$$

We will find the statistical characteristics of the signal component $\beta_h(\varepsilon)$ and noise component $\alpha_h(n_h)$ of the output effect of the discriminator.

Apparently, the DC bias is absent, as

$$\tilde{s}_{\lambda,h} = 0 \text{ и } \overline{z_{n,h}(n_{n,h})} = 0, n_0 = 0,$$

where n_0 is the constant component of the acting additive noise.

Note that here and further on, a wavy line above the variable means averaging over time, and a straight line means averaging over the set.

We will find the slope of the DC. In order to do it, we write down the time-averaged and set-averaged voltage of the processed (useful) signal at the output of the discriminator

$$\begin{aligned} \tilde{\beta}(\varepsilon) &= \left\{ \frac{1}{H} \sum_{h=1}^H s'_\lambda(\hat{\lambda}_h) m_1 \{ z'_{n,h,h} \} + \frac{1}{H} \sum_{h=1}^H s'_\lambda(\hat{\lambda}_{h-1}) \times \right. \\ &\times s'_\lambda(\hat{\lambda}_h) m_1 \{ z'_{n,h,h-1} \} + \frac{1}{H} \sum_{h=1}^H \left(s'_\lambda(\hat{\lambda}_{h-1}) \right)^2 m_1 \{ z'_{n,h-1,h-1} \} + \\ &+ \left. \frac{1}{H} \sum_{h=1}^H s'_\lambda(\hat{\lambda}_{h-1}) s'_\lambda(\hat{\lambda}_h) m_1 \{ z'_{n,h,h-1} \} \right\} \varepsilon. \end{aligned} \quad (3)$$

When $H \rightarrow \infty$

$$\lim \tilde{\beta}(\varepsilon) = \text{tr} E I_F \varepsilon = (E_0 I_{F.11} + 2E_1 I_{F.12} + E_0 I_{F.22}) \varepsilon,$$

where

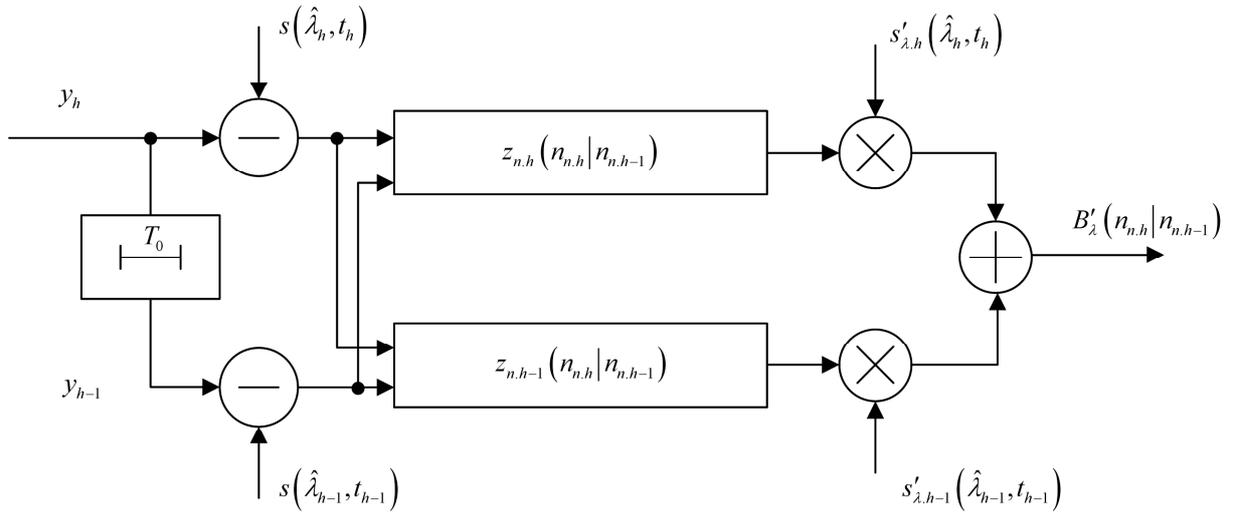


Fig. 1. Block diagram of the discriminator

$$E_0 = \lim_{H \rightarrow \infty} \frac{1}{H} \sum_{(h-1)=1}^H \left(s'_\lambda(\hat{\lambda}_{h-1}) \right)^2 =$$

$$= \lim_{H \rightarrow \infty} \frac{1}{H} \sum_{h=1}^H \left(s'_\lambda(\hat{\lambda}_h) \right)^2;$$

$$E_1 = \lim_{H \rightarrow \infty} \frac{1}{H} \sum_{h=1}^H s'_\lambda(\hat{\lambda}_{h-1}) s'_\lambda(\hat{\lambda}_h),$$

$I_{F,ij}$, $i, j = 1, 2$ are the elements of the Fisher information matrix:

$$I_{F,11} = -m_1 \left\{ \frac{\partial^2}{\partial n_{\Sigma,h-1}^2} \ln W_n(n_{\Sigma,h} | n_{\Sigma,h-1}) \right\};$$

$$I_{F,12} = -m_1 \left\{ \frac{\partial^2}{\partial n_{\Sigma,h} \partial n_{\Sigma,h-1}} \ln W_n(n_{\Sigma,h} | n_{\Sigma,h-1}) \right\};$$

$$I_{F,22} = -m_1 \left\{ \frac{\partial^2}{\partial n_{\Sigma,h}^2} \ln W_n(n_{\Sigma,h} | n_{\Sigma,h-1}) \right\}.$$

Here, a notation is introduced

$$r_s = \lim_{H \rightarrow \infty} \left[\frac{\sum_h s'_\lambda(\hat{\lambda}_{h-1}) s'_\lambda(\hat{\lambda}_h)}{\sum_h \left(s'_\lambda(\hat{\lambda}_h) \right)^2} \right].$$

We write the expression (3) as

$$\tilde{\beta}(\varepsilon) = \tilde{K}_{dc,c}^2 \varepsilon,$$

where $\tilde{K}_{dc,c}^2 \varepsilon = (I_{F,11} + 2r_s I_{F,12} + I_{F,22}) E_0$ is the slope of the DC under the influence of correlated additive noise.

We will find the characteristics of the fluctuation component of the discriminator.

Since the chains following the discriminator have response delay, the noise component in the bandwidth of an estimating device can be considered as «white noise» and characterized by its variance of output fluctuations of the discriminator:

$$\sigma_{fl}^2 = \tilde{\alpha}_h(n_n) =$$

$$= \left[s'_\lambda(\hat{\lambda}_h) z_{n,h}(n_{n,h}) + s'_\lambda(\hat{\lambda}_{h-1}) z_{n,h-1}(n_{n,h-1}) \right]^2 =$$

$$= \frac{1}{H} \left[\sum_h \left(s'_\lambda(\hat{\lambda}_h) \right)^2 m_1 \{ z_{n,h}^2 \} + 2 \sum_h s'_\lambda(\hat{\lambda}_h) s'_\lambda(\hat{\lambda}_{h-1}) \times \right.$$

$$\left. \times m_1 \{ z_{n,h} z_{n,h-1} \} + \sum_h \left(s'_\lambda(\hat{\lambda}_{h-1}) \right)^2 m_1 \{ z_{n,h-1}^2 \} \right].$$

As

$$m_1 \{ z_{n,h-1}^2 \} = m_1 \left\{ \left[\frac{\partial}{\partial n_{h-1}} \ln W_n(n_h | n_{h-1}) \right]^2 \right\} = I_{F,11}^n;$$

$$m_1 \{ z_{n,h} z_{n,h-1} \} =$$

$$= m_1 \left\{ \frac{\partial}{\partial n_h} \ln W_n(n_h | n_{h-1}) \frac{\partial}{\partial n_{h-1}} \ln W_n(n_h | n_{h-1}) \right\} = I_{F,12}^n;$$

$$m_1 \{ z_{n,h}^2 \} = m_1 \left\{ \left[\frac{\partial}{\partial n_h} \ln W_n(n_h | n_{h-1}) \right]^2 \right\} = I_{F,22}^n,$$

we obtain $\sigma_{fl}^2 = (I_{F,11}^n + 2r_s I_{F,12}^n + I_{F,22}^n) E_0$.

As an example, we consider the processing the deterministic signal $s(\lambda_h, t_h)$ exposed to additive correlated noise with lognormal PDF in the discriminator:

$$W_n(n_h | n_{h-1}) = \frac{\exp \left\{ -\frac{1}{2} f^2(n_h, n_{h-1}) \right\}}{n_h a \sigma_n^2 \sqrt{2\pi(1-r_n^2)}},$$

$$\text{where } f^2(n_h, n_{h-1}) = \frac{\left(\ln \frac{n_h}{c} - r_n \ln \frac{n_{h-1}}{c} \right)^2}{2a^2 \sigma_n^2 (1-r_n^2)};$$

a, c are coefficients.

Then, according to the above, we find

$$\tilde{K}_d^2 = \frac{1}{H} \sum_h \left\{ \frac{s'_\lambda(\hat{\lambda}_h)}{n_h} [1 + f^2(n_h, n_{h-1})] - r_n \frac{s'_\lambda(\hat{\lambda}_{h-1})}{n_{h-1}} [1 + f^2(n_h, n_{h-1})]^2 \right\}.$$

It can be shown that processing of a deterministic signal under the influence Gaussian correlated noise with the correlation coefficient r_n and the variance σ_n^2 :

$$\tilde{K}_d^2 = \rho K_d^2,$$

where \tilde{K}_d^2 is a slope of the discriminator DC, ρ is a coefficient characterizing the increase of the slope of the DC given the non-Gaussian nature of the PDF view $W_n(n_h | n_{h-1})$ of the influencing noise.

When $r_n = 1$

$$\rho = \frac{\exp\{2a^2\sigma_n^2\}}{(ac)^2};$$

$$\tilde{K}_d^2 = \frac{1}{H} \sum_{h=1}^H \frac{[s'_\lambda(\hat{\lambda}_h) - r_n s'_\lambda(\hat{\lambda}_{h-1})]^2}{\sigma_n^2 (1 - r_n^2)}.$$

When $r_n = 0$

$$\rho = \left(\frac{\exp\{2a^2\sigma_n^2\}}{ac^2} + \frac{\sigma_n^2 \exp\{2a^2\sigma_n^2\}}{c^2} \right).$$

Thus, if noise is described by logarithmically normal PDF, then, when the values of the coefficients a and c are given, it is easy to calculate the slope of the DC of the discriminator.

III. SYNTHESIS OF THE DISCRIMINATOR UNDER THE INFLUENCE OF ADDITIVE NOISE DESCRIBED BY MULTI-CONNECTED MARKOV SEQUENCES

Let the additive noise influencing the useful signal be represented by means of a k -bound stationary Markov sequence $\{n_{\Sigma, i-k}^i, i = -1, 0, 1, \dots\}$, which is fully characterized by the K -dimensional transition PDF $W_n(n_h | n_{h-1} \dots n_{h-k})$. We define the derivative of the LLF with a k -connected sequence as

$$B'_{\lambda, h} = \sum_{j=1}^k z_j(n_{\Sigma, h-k}^h) s'_\lambda(\hat{\lambda}_{h-j}), \quad j = \overline{1, k},$$

where

$$\begin{aligned} z_i(n_{\Sigma, h-k}^h) &= -\frac{\partial}{\partial n_{\Sigma, h-j}} \ln W_n(n_{\Sigma, h} | n_{\Sigma, h-k}^h) = \\ &= -\frac{\partial}{\partial n_{\Sigma, h-j}} \ln W_n(n_{\Sigma, h} | n_{\Sigma, h-1} \dots n_{\Sigma, h-k}). \end{aligned}$$

The functions z_j^i have the following properties:

$$m_1[z_k(n_{\Sigma, h}, \dots, n_{\Sigma, h-p})] = 0;$$

$$m_1[z_l(n_{\Sigma, i-k}^i) z_q(n_{\Sigma, i-k}^i)] = I_{F, lq}^\Sigma \delta_{ij},$$

(i.e. if $n_i \neq n_j$ the components of the vectors $z(n_i)$ and $z(n_j)$ are mutually uncorrelated), where $I_{F, lq}^\Sigma = m_1\{z_l(n_{i-k}^i) z_q(n_{i-k}^i)\}$ are the elements of the Fisher information matrix I_F^Σ .

Using the above results for the case of the k -bound stationary Markov sequence, we find

$$\tilde{\beta}(\varepsilon) = \text{tr}[Q I_{F, z}^\Sigma] \varepsilon = \sum_{i,j}^k Q_{ij} I_{F, jq}^\Sigma \varepsilon = \tilde{K}_d^2 \varepsilon,$$

where Q a positively defined matrix sized $k \times k$ with elements

$$Q_{ij} = \lim_{H \rightarrow \infty} \frac{1}{H} \sum_h s'_{\lambda, k} s'_{\lambda, j-i+k} Q_{j-i}; \quad Q_{aj} < \infty.$$

Then, for the variance of the fluctuation error

$$\sigma_{\beta}^2 = \text{tr}[Q I_{F, n}],$$

where $I_{F, n}$ is the Fisher information matrix of the influencing additive noise with elements

$$I_{F, n, ij} = m_1\{z_l(n_{\Sigma, i-k}^h) z_q(n_{\Sigma, i-k}^h)\}.$$

IV. ANALYSIS OF CHARACTERISTICS OF GENERALIZED DISCRIMINATOR

Having chosen extrapolated $\hat{\lambda}_{e, h}$ as a preliminary estimate of the information parameter so that $\hat{\lambda}_h^0 = \hat{\lambda}_{e, h}$, we write the nonlinear demodulation equations in non-Gaussian correlated noise using the observation model (1) as

$$\hat{\lambda}_h = \hat{\lambda}_{e, h} + \hat{\sigma}_{\lambda \lambda, h}^2 [B_{\lambda, h}^{n'} - F_{h, h-1} B_{\lambda, h-1}^{n'}]; \quad (4)$$

$$\hat{\sigma}_{\lambda \lambda, h}^2 = [B_{\lambda, h, h}^{\lambda'} + B_{\lambda, h, h}^{n'} - F_{h, h-1} (B_{\lambda, h, h-1}^{\lambda'} + B_{\lambda, h, h-1}^{n'})]^{-1}, \quad (5)$$

where $F_{h, h-1} = \frac{(B_{\lambda, h, h-1}^{\lambda'} + B_{\lambda, h, h-1}^{n'})}{(B_{\lambda, h-1, h-1}^{\lambda'} + B_{\lambda, h-1, h-1}^{n'} + \hat{\sigma}_{\lambda \lambda, h-1}^2)}$;

$$B_{\lambda, h}^{n'} = z_{n, h} s_{\lambda}''(\hat{\lambda}_{e, h}); \quad B_{\lambda, h-1}^{n'} = z_{n, h-1} s_{\lambda}'(\hat{\lambda}_{h-1});$$

$$B_{\lambda, h, h}^{n'} = B_{n, h, h}^{n'} [s_{\lambda}'(\hat{\lambda}_{e, h})]^2 - z_{n, h} s_{\lambda}''(\hat{\lambda}_{e, h});$$

$$B_{\lambda, h-1, h-1}^{n'} = B_{n, h-1, h-1}^{n'} [s_{\lambda}'(\hat{\lambda}_{h-1})]^2 - z_{n, h-1} s_{\lambda}''(\hat{\lambda}_{h-1});$$

$$B_{\lambda, h, h-1}^{n'} = B_{n, h, h-1}^{n'} s_{\lambda}'(\hat{\lambda}_{e, h}) s_{\lambda}'(\hat{\lambda}_{h-1}),$$

$$z_{n, h-i} = -\partial \ln W_n(n_h | n_{h-1}) / \partial \hat{n}_{h-i}; \quad \hat{n}_{h-i} = y_{h-i} - s(\hat{\lambda}_{h-i}); \quad i = 0, 1.$$

The equation (4) defines the algorithm for obtaining the optimal evaluation of the information process $\hat{\lambda}$, whereas (5) defines the development of the posterior variance.

At the step $h = 1$, the equations (4) and (5) take the form:

$$\hat{\lambda}_1 = \hat{\lambda}_{e, 1} + \hat{\sigma}_{\lambda \lambda, 1}^2 B_{\lambda, 1}^{n'}; \quad (6)$$

$$\hat{\sigma}_{\lambda \lambda, 1}^2 = [B_{\lambda, 1, 1}^{\lambda'} + B_{\lambda, 1, 1}^{n'}]^{-1}, \quad (7)$$

where $\hat{\lambda}_1$ is the initial quasi-optimal estimation of the information process λ , $\hat{\lambda}_{e,1}$ is the extrapolated value of the information process λ at the 1-st step, $\hat{\sigma}_{\lambda\lambda,1}^2$ is the posterior variance at the 1-st step, $\hat{\sigma}_{\lambda\lambda,1}^2$ is the second derivative of the logarithm of the transition PDF of the information process at the 1-st step, $B_{\lambda,1}^{n'}$ и $B_{\lambda,1}^{n''}$ are, respectively, the 1st and 2nd derivatives of the LLF with respect to the information parameter of the process at the 1-st step.

The equations (6) and (7) describe procedures in the demodulator for the case when the information sequence $\{\lambda_h\}$ and influencing additive $\{n_h\}$ non-Gaussian noise, which represent a process with independent values with the operator of species interaction with unambiguous inverse functions [6].

Expressions for the LLF under the influence of additive noise are determined from the expressions:

$$B_{\lambda,h-i}^{n'} \equiv \left. \frac{\partial \ln W_n(y_h - s(\lambda_h) | y_{h-1} - s(\lambda_{h-1}))}{\partial \lambda_{h-i}} \right|_{\substack{\lambda_h = \hat{\lambda}_{e,h} \\ \lambda_{h-1} = \hat{\lambda}_{h-1}}};$$

$$B_{\lambda,h-i,h-j}^{n''} \equiv \left. \frac{-\partial^2 \ln W_n(y_h - s(\lambda_h) | y_{h-1} - s(\lambda_{h-1}))}{\partial \lambda_{h-i} \partial \lambda_{h-j}} \right|_{\substack{\lambda_h = \hat{\lambda}_{e,h} \\ \lambda_{h-1} = \hat{\lambda}_{h-1}}},$$

$i, j = 0, 1$.

From the exact equations (4) and (5) under certain conditions we can proceed to simplified ones, in which second derivatives of the LLF are replaced by averaged values, and the posterior variance $\hat{\sigma}_{\lambda\lambda,h}^2 = \hat{\sigma}_{\lambda\lambda,h}^2 = \bar{\sigma}_{\varepsilon,h}^2$ is replaced by the averaged error variance $\bar{\sigma}_{\varepsilon,h}^2$.

As a result, we will obtain

$$\hat{\lambda}_h = \hat{\lambda}_{e,h} + \bar{\sigma}_{\varepsilon,h}^2 \left[z_{n,h} s'_{\lambda}(\hat{\lambda}_{e,h}) - \tilde{F}_{h,h-1} z_{n,h-1} s'_{\lambda}(\hat{\lambda}_{h-1}) \right];$$

$$\bar{\sigma}_{\varepsilon,h}^2 = \left[\bar{B}_{\lambda,h,h}^{\lambda'} + \tilde{B}_{\lambda,h,h}^{n''} - \tilde{F}_{h,h-1} \left(\bar{B}_{\lambda,h,h-1}^{\lambda'} + \tilde{B}_{\lambda,h,h-1}^{n''} \right) \right]^{-1};$$

$$\tilde{F}_{h,h-1} = \frac{\left(\bar{B}_{\lambda,h,h-1}^{\lambda'} + \tilde{B}_{\lambda,h,h-1}^{n''} \right)}{\left(\bar{B}_{\lambda,h-1,h-1}^{\lambda'} + \tilde{B}_{\lambda,h-1,h-1}^{n''} + \bar{\sigma}_{\varepsilon,h}^{-2} \right)},$$

where

$$\bar{B}_{\lambda,h-1,h-1}^{\lambda'} = -\frac{\partial^2 W_{\lambda}(\lambda_h | \lambda_{h-1})}{\partial \lambda_{h-1}^2} = I_{F,11}^{\lambda};$$

$$\bar{B}_{\lambda,h,h-1}^{\lambda'} = -\frac{\partial^2 W_{\lambda}(\lambda_h | \lambda_{h-1})}{\partial \lambda_{h-1} \partial \lambda_h} = I_{F,12}^{\lambda};$$

$$\bar{B}_{\lambda,h,h}^{\lambda'} = -\frac{\partial^2 W_{\lambda}(\lambda_h | \lambda_{h-1})}{\partial \lambda_h^2} = I_{F,22}^{\lambda};$$

$$\tilde{B}_{\lambda,h-1,h-1}^{n''} = \left(\tilde{s}'_{\lambda}(\hat{\lambda}_{h-1}) \right)^2 \frac{\partial^2 W_n(n_h | n_{h-1})}{\partial n_{h-1}^2} = P_{s'} I_{F,11}^n;$$

$$\tilde{B}_{\lambda,h,h-1}^{n''} = \left(\tilde{s}'_{\lambda}(\hat{\lambda}_{h-1}) \tilde{s}'_{\lambda}(\hat{\lambda}_h) \right)^2 \frac{\partial^2 W_n(n_h | n_{h-1})}{\partial n_{h-1}^2 \partial n_h} = R_{s'} I_{F,12}^n;$$

$$\tilde{B}_{\lambda,h,h}^{n''} = \left(\tilde{s}'_{\lambda}(\hat{\lambda}_h) \right)^2 \frac{\partial^2 W_n(n_h | n_{h-1})}{\partial n_h^2} = P_{s'} I_{F,22}^n;$$

$$P_{s'} = \lim_{H \rightarrow \infty} \frac{1}{H} \sum_{h=1}^H \left(s'_{\lambda}(\hat{\lambda}_{h-i}) \right)^2; i = 0, 1;$$

$$R_{s'} = \lim_{H \rightarrow \infty} \frac{1}{H} \sum_{h=1}^H \left(s'_{\lambda}(\hat{\lambda}_{h-1}) s'_{\lambda}(\hat{\lambda}_h) \right).$$

Here $I_{F,ij}^{\lambda}$ и $I_{F,ij}^n$ are the elements of the Fisher information matrix of the information process $\{\lambda_h\}$ and additive noise $\{n_h\}$ respectively.

By setting

$$r_{s'} = \lim_{H \rightarrow \infty} \frac{1}{H} \left[\frac{\sum_{h=1}^H \left(s'_{\lambda}(\hat{\lambda}_{h-1}) s'_{\lambda}(\hat{\lambda}_h) \right)}{\sum_{h=1}^H \left(s'_{\lambda}(\hat{\lambda}_{h-1}) \right)^2} \right],$$

we can show that $R_{s'} = r_{s'} P_{s'}$.

By using the obtained relations in the demodulation algorithm, we obtain:

$$\hat{\lambda}_h = \hat{\lambda}_{e,h} + \bar{\sigma}_{\varepsilon,h}^2 \left[z_{n,h} s'_{\lambda}(\hat{\lambda}_{e,h}) + \tilde{F} z_{n,h-1} s'_{\lambda}(\hat{\lambda}_{h-1}) \right];$$

$$\bar{\sigma}_{\varepsilon,h}^2 = \left[I_{F,22}^{\lambda} + P_{s'} I_{F,22}^n + \tilde{F} \left(I_{F,12}^{\lambda} + r_{s'} P_{s'} I_{F,12}^n \right) \right]^{-1};$$

$$\tilde{F} = -\frac{\left(I_{F,12}^{\lambda} + R_{s'} I_{F,12}^n \right)}{\left(I_{F,11}^{\lambda} + P_{s'} I_{F,11}^n + \bar{\sigma}_{\varepsilon,h}^{-2} \right)}.$$

In this case, the discriminatory characteristic is understood as

$$D = \left[z_{n,h} s'_{\lambda}(\hat{\lambda}_{e,h}) + \tilde{F} z_{n,h-1} s'_{\lambda}(\hat{\lambda}_{h-1}) \right].$$

The above results are true for the analysis of the characteristics of such a discriminator, thus it can be shown that

$$D = \alpha_{d,h}(n_h) + \beta_{d,h}(\varepsilon),$$

where $\alpha_{d,h}(n_h)$, $\beta_{d,h}(\varepsilon)$ are, respectively, the noise discriminator function and the signal discriminator function;

$$\alpha_{d,h}(n_h) = z_{n,h} s'_{\lambda}(\hat{\lambda}_{e,h}) + \tilde{F} z_{n,h-1} s'_{\lambda}(\hat{\lambda}_{h-1});$$

$$\beta_{d,h}(\varepsilon) = \varepsilon \left\{ \left(s'_{\lambda}(\hat{\lambda}_{e,h}) \right)^2 z'_{n,h,h} + s'_{\lambda}(\hat{\lambda}_{e,h}) s'_{\lambda}(\hat{\lambda}_{e,h-1}) z'_{n,h,h-1} + \tilde{F} \left[\left(s'_{\lambda}(\hat{\lambda}_{h-1}) \right)^2 z'_{n,h-1,h-1} + s'_{\lambda}(\hat{\lambda}_{h-1}) s'_{\lambda}(\hat{\lambda}_{e,h}) z'_{n,h,h-1} \right] \right\}.$$

Consequently, the DC is

$$\tilde{\beta}_{d,h}(\varepsilon) = \tilde{K}_d^2 \varepsilon,$$

where $\tilde{K}_d^2 = \left(I_{F,11}^n \tilde{F} + r_{s'} \left(1 + \tilde{F} \right) I_{F,12}^n + I_{F,22}^n \right) E_0$.

In this case, the variance of the fluctuation component will be described by the equation

$$\begin{aligned}\sigma_{d,\beta}^2 &= \tilde{\alpha}_{d,h}(n_n) = \left[z_{n,h} s'_\lambda(\hat{\lambda}_{e,h}) + z_{n,h-1} s'_\lambda(\hat{\lambda}_{h-1}) \tilde{F} \right] = \\ &= \left[I_{F,11}^n \tilde{F} + 2r_s I_{F,12}^n \tilde{F} + I_{F,22}^n \right] E_0.\end{aligned}$$

It should be noted that if $\tilde{F} = 1$, the characteristics of a generalized discriminator coincide with the characteristics defined above.

V. ANALYSIS OF THE EFFICIENCY OF THE DISCRIMINATOR IN THE CONDITIONS OF BIASED HYPOTHESES

Under the conditions of biased hypotheses, we will analyze the characteristics of the discriminator. In this case, the sample $\{n_{n,\Sigma}\}$ belongs to the PDF $U(n_{n,\Sigma})$, whereas the structure of the discriminator is set and calculated for the PDF $W_n(n_{n,\Sigma})$.

We will analyze the noise with independent values on the basis of simplified calculations, but with the preservation of all the results.

In this case, the structure of the discriminator will be described by the LLF as follows

$$B'_\lambda = s'_\lambda(\hat{\lambda}_{e,h}) z(n_{n,\Sigma}),$$

where $z(n_{n,\Sigma}) = -\frac{d}{dn_{n,\Sigma}} \ln W_n(n_{n,\Sigma})$.

We determine the average value of the statistics at the discriminator output when the hypothesis is biased and when $n_\Sigma \in U(n_{n,\Sigma})$:

$$\begin{aligned}\tilde{B}'_\lambda &= \frac{1}{H} \sum_h s'_\lambda(\hat{\lambda}_{e,h}) m_1 \{z(n_{n,\Sigma}) | n_{n,\Sigma} \in U(n_{n,\Sigma})\} = \\ &= \frac{1}{H} \sum_h s'_\lambda(\hat{\lambda}_{e,h}) \int_{-\infty}^{+\infty} z(n_{n,\Sigma}) \mathcal{U}(n_{n,\Sigma}) d(n_{n,\Sigma}).\end{aligned}$$

Here, the DC will be described by the expression

$$\tilde{\beta}(\varepsilon)^* = \frac{1}{H} \sum_h (s'_\lambda(\hat{\lambda}_{e,h}))^2 I_{F,WU} \varepsilon = \tilde{K}_d^{*2} \varepsilon,$$

where $I_{F,WU} = \int_{-\infty}^{+\infty} z(n_{n,\Sigma}) U(n_{n,\Sigma}) d(n_{n,\Sigma})$.

The variance of the fluctuation component, is described by the expression

$$\sigma_{d,\beta}^{2*} = \frac{1}{H} \sum_h s'_\lambda(\hat{\lambda}_{e,h}) M_2 \{z(n_{n,h}) | U(n_{n,h})\} = E_s I_{WU},$$

where

$$I_{WU} = \int_{-\infty}^{+\infty} z^2(n_n) U(n_n) d(n_n) - \left(\int_{-\infty}^{+\infty} z(n_n) U(n_n) d(n_n) \right)^2.$$

The analysis of the discriminator's work in the conditions of biased hypotheses showed that in this case what has to be done is just replace I_{F,n_Σ} for $I_{F,WU}$, and $I_{F,n}$ for I_{WU} .

VI. CONCLUSIONS

Thus, we carried out the analysis of discriminators of tracking meters measuring information parameters of the signal under the influence of additive correlated non-

Gaussian noise described by single-connected and multi-connected Markov sequences. The structural diagrams of the discriminator operating under the influence of correlated Gaussian noise and additive non-Gaussian noise are obtained. It is shown if noise is non-Gaussian, it decreases due to both decorrelation and nonlinear amplitude suppression. Generally, a linear decorrelator of noise should be coordinated with the spectral density of the power of influencing noise. Main characteristics of the discriminator are considered and analyzed.

An example of a deterministic signal processed under the influence of additive correlated noise with a logarithmically normal PDF is given.

The analysis of characteristics of the generalized discriminator is carried out. The efficiency of the discriminator's work in the conditions of biased hypotheses is analyzed.

REFERENCES

- [1] Van Trees, K. Bell, and Z. Tiany, Detection Estimation and Modulation Theory, 2nd Edition, Part I, Detection, Estimation, and Filtering Theory. London: Wiley & Sons, Inc., 2013.
- [2] V. P. Tuzlukov, Signal Processing Noise, Boca Raton, London, New York, Washington D.C.: CRC Press, Taylor & Francis Group, 2002.
- [3] Ellingson, Radio System Engineering. Cambridge University Press, 2016.
- [4] M. Barkat, Signal Detection and Estimation. Norwood: Artech House, 2005.
- [5] H. S. Kassam, Signal Detection in Non-Gaussian Noise. Berlin: Springer, 1988.
- [6] V. M. Artyushenko, and V. I. Volovach, "Synthesis and analysis of discriminators under influence broadband non-Gaussian noise", IOP Conference Series XI International scientific and technical conference "Applied Mechanics and Dynamics Systems", J. Phys.: Conf. Ser., January 2018, Vol. 944, No. 1, 012004. DOI: [10.1088/1742-6596/944/1/012004](https://doi.org/10.1088/1742-6596/944/1/012004)
- [7] J. Yang, Y. Cheng, H. Wang, Y. Li, and X. Hua, "Unknown stochastic signal detection via non-Gaussian noise modeling," Proceedings 2015 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Ningbo, China, 2015, pp. 1–4. DOI: [10.1109/ICSPCC.2015.7338861](https://doi.org/10.1109/ICSPCC.2015.7338861)
- [8] E. Palahina, and V. Palahin, "Signal detection in additive-multiplicative non-Gaussian noise using higher order statistics", 2016 26th International Conference Radioelektronika (RADIOELEKTRONIKA), IEEE, 2016, pp. 262-267. DOI: [10.1109/RADIOELEK.2016.7477367](https://doi.org/10.1109/RADIOELEK.2016.7477367)
- [9] V. M. Artyushenko, V. I. Volovach, and V. N. Budilov, "Synthesis and analysis of discriminators meter information parameters signal under non-Gaussian noise with band-limited spectrum", Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2017). Novi Sad, Serbia, Sept 29-Oct 2, 2017. – Kharkov: KNURE, 2017. P. 355-358. DOI: [10.1109/EWDTS.2017.8110112](https://doi.org/10.1109/EWDTS.2017.8110112)
- [10] V. M. Artyushenko, and V. I. Volovach, "Adaptive signal processing nonlinear block with quadrature generators under influence broadband noise", Proceedings of 2018 IEEE East-West Design & Test Symposium (EWDTS'2018). Kazan, Russia, Sept 14-17, 2018. DOI: [10.1109/EWDTS.2018.8524710](https://doi.org/10.1109/EWDTS.2018.8524710)
- [11] V. M. Artyushenko, and V. I. Volovach, "Measuring information signal parameters under additive non-Gaussian correlated noise", Optoelectronics, Instrumentation and Data Processing, 2016, Vol. 59, No. 6, pp. 22-28. DOI: [10.15372/AUT20160603](https://doi.org/10.15372/AUT20160603)
- [12] V. M. Artyushenko, and V. I. Volovach, "Adaptive signal processing nonlinear block with quadrature generators under influence broadband noise", Proceedings of 2018 IEEE East-West Design & Test Symposium (EWDTS'2018). Kazan, Russia, Sept 14 – 17, 2018. DOI: [10.1109/EWDTS.2018.8524710](https://doi.org/10.1109/EWDTS.2018.8524710)

Accurate Soft Error Rate Reduction using Modified Resolution Method

Alexander Stempkovskiy
DSc, prof., scientific director at Institute
for Design Problems in
Microelectronics
Moscow, Russia
stal09@ippm.ru

Dmitry Telpukhov
PhD, head of department at
Institute for Design
Problems in
Microelectronics
Moscow, Russia
NoFrost@inbox.ru

Vladislav Nadolenko
Design engineer
at Institute for Design
Problems in
Microelectronics
Moscow, Russia
vl777nd@list.ru

Abstract—The influence of cosmic radiation on integrated circuits is one of the most important reliability challenges of modern semiconductor technologies, and the most significant problems are associated with single event effects. Consequently, even at the early stage of the logical synthesis of combinational circuits, it is necessary to take into account the requirements for the reliability of operation. The article proposes an iterative method of resynthesis of combinational circuits by using implicit don't cares. The extraction procedure of these don't cares is performed using a modified resolution method, which has been widely used in the field of automated reasoning and especially in automated theorem proving. The applicability of the iterative process is ensured by effective metrics and methods for evaluating the reliability of combinational circuits. The proposed approach differs from the known methods in the absence of errors in finding logical constraints and accurate estimation of the soft error rate (SER) at each iteration. The effectiveness of the method is demonstrated on a large set of benchmark circuits.

Keywords—soft errors, resynthesis, fault tolerance, observability, ODC, Modified Resolution Method.

1. INTRODUCTION

Under influence of the harsh environmental conditions, various radiation effects arise in electronic devices. The main focus is currently on Single Event Effects (SEE) and Total Ionizing Dose (TID). These effects can lead to both a short-term malfunction of the equipment — an occasional soft error, or a complete hardware failure — a hard error. Nowadays the influence of the total ionizing dose effects decreases by the advance of technology process and is mainly solved by different radiation hardening methods. In contrast, single effects start to play pivotal role, and it is now necessary to apply special approaches in SEE mitigation circuit design.

All single effects, in turn, are divided into soft errors in memory cells and soft errors in the combinational parts of integrated circuits. Historically, memory elements were considered the most vulnerable to the occurrence of error, but in recent years, as a result of many different technological trends, the frequency of occurrence of soft errors in logical devices often exceeds the frequency of failures in sequential elements and memory cells [1][2].

In many previous works, it was shown that even at the early stages of logical design it is possible to implement circuits with

the SEE mitigation property which improves circuit's reliability. In general, methods for such fault-tolerant circuits design are based on the mechanism of logical masking. This mechanism is inherent in all combinational circuits and lies in the fact that failures that occur at the logic elements at a particular point, do not affect the primary outputs of the circuit. It was shown that different circuits that perform the same function can differ significantly in this property. In our work, we propose to improve the logical masking property of combinational circuits by finding logical constraints that allow adding small patches that protect certain parts of the circuit.

2. PREVIOUS WORK

A large volume of articles is devoted to the development of metrics and methods for assessing the fault-tolerance of combinational circuits. Part of the work is devoted to the development of metrics that link various masking mechanisms into one metric [3]. In other papers, only the logical masking is considered, and the main efforts are aimed at finding the best compromise between speed and accuracy loss, that arise due to the presence of re-convergent paths [4]. Classical approaches in this area are methods based on probabilistic transfer matrices (PTM) [5] and various probabilistic approaches [6],[7]. An efficient approach based on back propagation was developed for use in iterative algorithms that are critical to the accuracy of the observability masks [8].

The main part of the work is devoted to methods for improving the fault tolerance of combinational circuits. All methods can be divided by the criterion of systemacy of adding redundancy. Redundancy is systematically introduced in methods such as TMR[9], quadded logic [10], concurrent error detection [11][12]. In these approaches, redundancy is introduced in a strictly defined, predictable way, which makes it possible to predict efficiency and overhead in advance. Another way to introduce redundancy is the approach associated with logic resynthesis [13][14]. This class of methods uses an iterative approach, where within each iteration a vulnerable section of the circuit locally changes. To apply such an iterative approach, it is necessary to determine the procedure for selecting a place for a local change, the substitution procedure itself and the goal function. In

comparison with the systematic introduction of redundancy, these methods have greater flexibility, less overhead, but have less predictability, since they often rely on heuristic algorithms.

The method proposed in this paper is based on another iterative approach that uses logical implications in circuit to introduce redundancy [13]. The key difference lies in accuracy of resynthesis. We use Modified Resolution Method (MRM) [15] to find valid logical constraints in circuit and then extract implications. Other options are tracking implication paths and using probabilistic methods [13]. The former is less effective than MRM, and the latter has non-zero probability of yielding incorrect result which may lead to change in circuit functionality during resynthesis.

3. MODIFIED RESOLUTION METHOD

The key definition for further reasoning is the notion of a logical constraint. By the logical constraint in the combinational circuit, we mean the combination of logical states at the nodes of the circuit that are impossible for any values on primary inputs. For example, consider a simple circuit of one element AND2 (Fig 1).

The function of this element is shown in the first truth table. The second table defines the function $(a \& b) \leftrightarrow y$, which determines the mutual states of the nodes that are possible for this circuit. The rows of this table, corresponding to the zeros of the function, define logical constraints that we will write as elementary conjunctions. Thus, all the logical constraints for this circuit are as follows: $\{\bar{a}by, \bar{a}b\bar{y}, a\bar{b}y, a\bar{b}\bar{y}\}$.

For each logical element with N inputs, 2^N primary restrictions can always be formed. However, for large circuits, there are more complex constraints that can bind several different elements. To find them, one can use the effective symbolic method based on the resolution method. The full set of logical constraints in circuit forms its SDC (Satisfiability Don't Care).

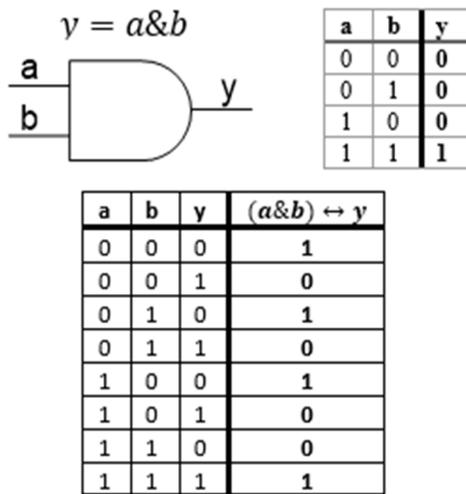


Figure 1. Logical Constraints extraction for and2 gate

The resolution method is a method of deriving new Boolean relations from a given set of Boolean relations. In its original form, the resolution rule is as follows: The result of the $X + F$ and $\bar{X} + G$ clauses is the $F + G$ clause. In other words:

$$(X + F) \cdot (\bar{X} + G) \rightarrow F + G \quad (1)$$

For the problem of finding new logical constraints, this method cannot be directly used. Transform equation (1) according to de Morgan's rule:

$$(\bar{X} \cdot \bar{F}) \cdot (\bar{X} \cdot \bar{G}) \rightarrow \bar{F} \cdot \bar{G} \quad (2)$$

Next, we use the fact that the negation of variable is equal to its equivalence operation with zero: $\bar{a} = a \leftrightarrow 0$. We obtain:

$$((\bar{X} \cdot \bar{F}) \leftrightarrow 0) \cdot ((\bar{X} \cdot \bar{G}) \leftrightarrow 0) \rightarrow ((\bar{F} \cdot \bar{G}) \leftrightarrow 0)$$

Making the change of variables $\bar{G} = A$ and $\bar{F} = B$, we get the final expression for the modified resolution method:

$$((X \cdot A) \leftrightarrow 0) \cdot ((\bar{X} \cdot B) \leftrightarrow 0) \rightarrow ((A \cdot B) \leftrightarrow 0) \quad (3)$$

Its essence lies in the fact that since the expressions $X \cdot A$ and $\bar{X} \cdot B$ are equal to zero, it follows that the expression $A \cdot B$ is zero. Two elementary conjunctions, which are zero, generate a new conjunction, which is also zero. Based on Fig. 1, it follows that this rule is suitable for generating new logical constraints from existing ones.

It is easy to show that each logical constraint containing exactly two literals generates two implication relations between those literals. Consider, for example, constraint $A \cdot B$. When A has value of 1, B is forbidden to have value of 1. Therefore, relation $A \rightarrow \bar{B}$ is always valid, i.e. $A \rightarrow \bar{B} \equiv 1$. Similarly, when B is 1, A has to be 0: $B \rightarrow \bar{A} \equiv 1$. This principle allows us to search for valid implication relations in circuit using logical constraints and MRM.

4. FAST SER ESTIMATION PRINCIPLES

In previous research we used logic sensitivity factor of a circuit [4] as a fault-tolerance metric. Sensitivity factor characterizes logic masking features of circuit and is calculated by formula:

$$sf = \sum_{G_i} (O_{G_i}) \quad (4)$$

where O_{G_i} is probability of observability of gate G_i at primary circuit outputs, summation over all gates takes place.

Currently we also use another metric – circuit's "sensitive area":

$$sa = \sum_{G_i} (O_{G_i} * A_{G_i}) \quad (5)$$

where A_{G_i} is area of gate G_i .

Take note that given metric is not equal to area of all sensitive parts in the circuit. It merely considers probabilities of each gate being hit by particle.

In this research, bit-parallel simulation and vector calculations [16] are used to get statistics which is further used to calculate O_{G_i} for each gate. Signature is assigned to each circuit node:

$$Signature_{G_i} = \{f_{G_i}(X_1), f_{G_i}(X_2), \dots, f_{G_i}(X_N)\} \quad (6)$$

where X_j is the j-th set of input signals, N is the number of simulated input vectors, $f_{G_i}(X)$ is logic function of node G_i with respect to circuit primary inputs. For convenience, only gates

with one output are considered, and therefore each circuit node uniquely corresponds to either gate or primary input that controls it. However, these considerations can be easily generalized to multiple-output gates.

Node observabilities are presented by ODC (Observability Don't Care) masks [17]:

$$ODC_{G_i} = \{O_{G_i}(X_1), O_{G_i}(X_2), \dots, O_{G_i}(X_N)\} \quad (7)$$

where $O_{G_i}(X)$ is observability of gate G_i at primary circuit outputs with respect to input vector X . Each possible input vector determines the corresponding state of a given combinational circuit, i.e. logic values of all nodes in the circuit. And in each possible state gate G_i is either observable ($O_{G_i}(X) = 1$) or not observable ($O_{G_i}(X) = 0$) at circuit primary outputs. Probability of observability O_{G_i} can be obtained as ratio of number of 1's in ODC_{G_i} vector (Hamming's weight) to its length:

$$O_{G_i} = \frac{w(ODC_{G_i})}{|ODC_{G_i}|} \quad (8)$$

Signatures and ODC masks are stored in memory as integer variables and calculations are performed using bitwise operations.

First, signatures of primary circuit inputs are generated according to the given set of input vectors. Then signatures of all gates are calculated in topological order. Gate's logic function with bitwise operators is used to propagate signatures from its inputs to its output.

ODC masks are more difficult to calculate. There are several methods for obtaining node observabilities in particular circuit states [17,18]. The fastest of them, the method of back propagation of ODC masks, has linear time complexity with respect to the number of gates. However, it is not exact if there are reconvergent paths in the circuit [18]. Downstream error simulation provides precise result but has quadratic time complexity in the number of gates. Modular approach [6] can be used as a trade-off between the time and precision. Small parts of circuit containing reconvergent paths are represented as modules. Error simulation is used inside modules to compute ODC masks of their internal gates while at the top level of hierarchy back propagation method is used. This method proved to improve precision compared to back propagation with acceptable time cost. But in some specific cases complex analysis of the circuit structure may be necessary to approach the precise result. In our work, we use method of accelerated error simulation, developed earlier [8]. Timing costs are significantly higher for this approach, but quality of resynthesis depends on the accuracy of ODC masks calculation significantly. It is showed in section 6 that linear-time back propagation method is highly unreliable during implication-based resynthesis.

5. ITERATIVE ALGORITHM FOR SER REDUCTION

Our resynthesis algorithm includes two stages: searching for implications and adding redundant patches.

First stage starts with extracting each gate's logic constraints as shown at Fig. 1 and reducing them to minimal set. Resolution operation (3) is useless when applied to logical constraints extracted from one gate – its result is always included in current

minimal set. However, internal circuit nodes are adjacent to more than one gate. Logical constraints from different initial subsets may be successfully used as inputs to resolution operation (Fig. 2). Newly created logical constraints at Fig. 2 originate from both NAND2 and INV elements and cannot be crossed with any of their initial constraints. In general, resolution operation usage is limited to pairs of constraints with non-intersecting originitive sets of gates.

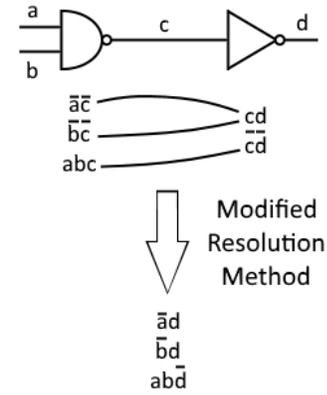


Figure 2. Example of resolution operation

Basic algorithm for searching logic constraints with modified resolution method is as follows:

```

ConstrSet = ∅
for each gate in circuit:
    extract subset of constraints from logic function
    minimize subset
    subset → ConstrSet
repeat N times:
    resolved = ∅
    for each pair in ConstrSet × ConstrSet:
        if possible:
            resolve(pair) → resolved
    minimize resolved
    resolved → ConstrSet

```

Figure 3. Basic procedure of searching logic constraints with modified resolution method

However, amount of logical constraints in circuit is exponential in its size in gates, which makes exhaustive search impossible even for relatively small circuits (hundreds of gates). To address this issue and improve scalability, we add limitations to the algorithm.

First, implications originate from logic constraints with two literals only. Thus, constraints with higher length are useless for the next stage. However, each resolution operation can reduce constraint length by one. For example, $a\bar{b}y$ and $\bar{a}by$ collapse into $\bar{b}y$. In this case resolution operation is equivalent to minterm merging. Potentially useful logical constraints are therefore limited to those with length less or equal to $2 + M$ where M is number of iterations remaining.

Second, implications are unlikely to bind too distant nodes in circuit. Close nodes, however, are generally less effective for resynthesis. Distance between nodes in output logical constraints is limited at 2^N where N is number of iterations. By distance we mean number of gates in shortest path from one node to another. That path is not necessarily valid as a signal path – nodes bound by implication can be connected through shared fan-in cone. Considering that implications binding nodes at distance more than 8 are rare, we choose $N = 3$ for our algorithm.

These limitations, although practical, make precise algorithm complexity assessment impossible. Generally, the more branches circuit's wires have, the more time-consuming constraint search may become.

Next stage is essentially an exhaustive search through implications found at previous step. Each implication is a basis for possible observability-reducing patch (Fig. 4). For further explanation we use following definitions:

- T – target node, the one that is being protected;
- S – source node, auxiliary input for patch;
- T^* – patch output replacing node T in circuit.

Logical functions of patches $T^*(T, S)$ are known for all four implication types ($S \rightarrow T, S \rightarrow \bar{T}, \bar{S} \rightarrow T, \bar{S} \rightarrow \bar{T}$). For example, if $S \rightarrow T \equiv 1$ then OR2 element serves as a patch: $T^* = T + S$. Indeed, if $S = 1$ implies $T = 1$ then it can be a controlling value for $T^* = 1$. And when $S = 0$, T^* must be equal to T. OR2 element meets those requirements. Similar reasoning can be done for other three implication types. However, there is also another way. Further we present method that allows creating patch based on arbitrary logical constraint thus generalizing implication-based resynthesis to SDC-based resynthesis.

In general, observability-reducing patch must meet two requirements:

- $T^* = T$ for all possible input combinations;
- T^* function depends on S (or each S_i in set of sources for arbitrary logical constraint).

Meeting both requirements simultaneously is possible due to SDC on patch inputs. Consider logical constraint $L(T, S_1, \dots, S_n)$ with $n + 1$ literals. The following function is suited for patch:

$$T^*(T, S_1, \dots, S_n) = T \oplus L(T, S_1, \dots, S_n) \quad (9)$$

The first example with implication $S \rightarrow T$ is then transformed as follows. Implication is replaced with corresponding logical constraint $S\bar{T}$. Patch function T^* is then derived from (9): $T^* = T \oplus S\bar{T} = T + S$.

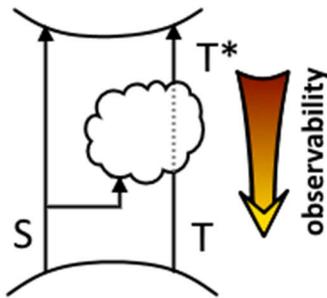


Figure 4. Scheme of implication-based patch

```

implications = ∅
for each constraint in ConstrSet
  if length(constraint) == 2:
    derive implications
GF0 = goal_function(circuit)
for each implication
  add corresponding patch to circuit
  GF1 = goal_function(circuit)
  if GF1 > GF0:
    GF0 = GF1
  else:
    remove patch
  minimize resolved
  resolved → ConstrSet

```

Figure 5. Basic algorithm for circuit resynthesis with implication-based patches

Each potential patch can reduce observability of node T and its fan-in cone and at the same time increase observability of node

S and its fan-in cone. Moreover, patch itself is prone to errors. Therefore, not every patch is effective and goal function must be evaluated at each iteration of resynthesis. The main algorithm of searching and applying effective patches is shown at Fig. 5. Arbitrary logical constraints are not used in current experiments.

Time complexity of this algorithm is $O(|C_2| \cdot N^2)$, where $|C_2|$ is cardinality of set of logical constraints of length 2, N is number of gates in circuit. Efficient heuristics are needed to decrease both multipliers.

6. EXPERIMENTAL RESULTS

ISCAS'85 and LGSynth'89 benchmark circuits were used as input for experiments. Limited SDC search allowed us to process circuits of up to thousands of gates size. However, further scaling is rather impossible before the faster methods for patch efficiency evaluation are developed.

Results in table 1 show that SER reduction and area overhead vary widely from one circuit to another. Also, it can be seen that the proposed method alone cannot affect circuit's reliability dramatically as sensitive area reduction rarely reaches 15%. However, there are two advantages as well.

First is low area overhead for particular circuits. Cordic and i4 from LGSynth'89 benchmark have $-\Delta A / \Delta SA < 1$. Some other circuits have $-\Delta A / \Delta SA \approx 2$. For TMR this coefficient is more than 3. High area overhead on circuits c880, count and tt2 can be explained by the fact that goal function addressed only reliability.

Second advantage of the proposed method is compatibility with other resynthesis methods. Table 2 represents results of the same experiment on circuits modified by our local rewriting algorithm [14]. Last column contains relative sensitive area reduction after both methods. Results show that different techniques work better on different circuits and their combination can be used for creating more efficient and stable algorithm.

TABLE I. RESULTS OF IMPLICATION-BASED RESYNTHESIS FOR BENCHMARK CIRCUITS

Benchmark	Circuit	Gates	Implications	Patches applied	ΔSA , %	ΔA , %
ISCAS'85	c1908	499	4410	72	-6.49%	17.59%
	c499	306	2034	11	-2.18%	4.00%
	c880	313	973	45	-4.60%	18.04%
LGSynth'89	cm151a	23	54	2	-7.07%	14.29%
	cordic	45	28	3	-13.06%	7.87%
	count	176	2410	51	-5.65%	39.38%
	frg2	615	7654	200	-15.90%	41.99%
	i4	146	41	6	-6.74%	5.75%
	term1	162	904	41	-19.63%	32.06%
	tft2	180	1248	38	-6.67%	26.73%

TABLE II. RESULTS OF IMPLICATION-BASED RESYNTHESIS FOR CIRCUITS AFTER LOCAL REWRITING

Benchmark	Circuit	Gates	Implications	Patches applied	ΔSA , %	ΔA , %	ΔSA (overall), %
ISCAS'85	c1908	507	3521	42	-3.25%	9.18%	-18.21%
	c499	334	2360	19	-1.18%	6.49%	-16.41%
	c880	309	687	21	-1.66%	7.67%	-8.60%
LGSynth'89	cm151a	19	23	1	-12.48%	5.33%	-16.07%
	cordic	51	33	2	-2.29%	4.81%	-14.53%
	count	150	1092	43	-14.37%	33.55%	-21.95%
	frg2	586	4477	179	-14.23%	36.68%	-23.76%
	term1	170	921	33	-18.49%	22.90%	-22.41%
	tft2	174	1034	46	-9.48%	30.19%	-22.68%

TABLE III. RESULTS OF IMPLICATION-BASED RESYNTHESIS WITH BACK PROPAGATION METHOD

Circuit	Initial back propagation error, %	Back propagation error after resynthesis, %	ΔSA (evaluated using back propagation), %	ΔSA (exact), %
c2670	0.51%	-24.11%	-18.68%	7.70%
c499	3.50%	-11.96%	-14.30%	0.75%
c5315	1.25%	-9.44%	-8.30%	2.52%
c7552	-0.06%	-6.80%	-6.10%	0.69%
c880	1.78%	-13.21%	-13.99%	0.88%

The last experiment was conducted in order to test reliability of back propagation algorithm for gate observability computation and evaluation of circuit's sensitive area. It was used to guide the resynthesis and then the result was checked with exact method. As table 3 shows, back propagation method appeared unreliable to the extent of reducing circuit's sensitive area. The reason is that patches always bring more re-convergent paths to circuit. The more patches are added, the less accurate back propagation method becomes.

7. CONCLUSION

In this work, we developed an algorithm for implication-based resynthesis aimed at improving circuit fault-tolerance, which is accurate in two different ways. First, its output circuit is always functionally equivalent to the input circuit due to correct implication search with modified resolution method. Second, re-convergent paths are considered during logical

masking probability evaluation. Experiment results confirmed that they cannot be ignored in context of implication-based resynthesis. We also showed that our algorithm can improve logical masking in circuit, reducing its SER. Preliminary use of local rewriting can further increase circuit fault-tolerance.

Finally, we proposed new resynthesis method based on logical constraints and showed that implications are essentially a subset of circuit's SDC. Further work will be aimed at developing time-saving heuristics for patch efficiency evaluation. Faster computation methods will allow us to implement SDC-based resynthesis effectively.

REFERENCES

- [1] T. Yaran, S. Tosun Improving combinational circuit resilience against soft errors via selective resource allocation // 2017 IEEE 20th International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS)
- [2] Mahatme NN, Gaspard NJ, Assis T, Jagannathan S, Chatterjee I, Loveless TD, et al. Impact of technology scaling on the combinational

- logic soft error rate. In: International reliability physics symposium (IRPS); 2014. p. 5F.2.1–5F.2.6.
- [3] M. Raji, H. Pedram, B. Ghavami, "A practical metric for soft error vulnerability analysis of combinational circuits", *Microelectron. Rel.*, vol. 55, no. 2, pp. 448-460, 2015.
- [4] Stempkovskiy, A., Telpukhov, D., Solovyev, R., Balaka, E., Naviner, L. Practical metrics for evaluation of fault-tolerant logic design // Proceedings of the 2017 IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017, pp. 569-573
- [5] Krishnaswamy S, Viamontes GF, Markov IL, Hayes JP. Probabilistic transfer matrices in symbolic reliability analysis of logic circuits. *ACM Trans Des Automation Electr Syst* 2008;13(1).
- [6] Han J, Chen H, Boykin E, Fortes J. Reliability evaluation of logic circuits using probabilistic gate models. *Microelectronics Reliability* 2011;51(2):468–76.
- [7] V.H. Vaghef A. Peiravi, Node-to-node error sensitivity analysis using a graph based approach for VLSI logic circuits *Microelectronics Reliability* 2015;55: 264–271
- [8] Stempkovskiy, A., Telpukhov, D., Nadolenko, V. Fast and accurate back propagation method for reliability evaluation of logic circuits // Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018 pp. 1424-1429
- [9] W.H. Pierce, *Failure-Tolerant Computer Design*, Academic Press, USA, 1965.
- [10] J.G. Tryon, "Quadded Logic," *Redundancy Techniques for Computing Systems*, 1962, pp. 205-228.
- [11] Gavrilov, S.V., Gurov, S.I., Zhukova, T.D., Rukhlov, V.S., Ryzhova, D.I., Tel'pukhov, D.V. Methods to Increase Fault Tolerance of Combinational Integrated Microcircuits by Redundancy Coding // *Computational Mathematics and Modeling*, 28(3), pp. 400-406, 2017
- [12] Stempkovskiy, A., Telpukhov, D., Gurov, S., Zhukova, T., Demeneva, A. R-code for concurrent error detection and correction in the logic circuits // Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, (ElConRus 2018), pp. 1430-1433
- [13] S. Krishnaswamy, S. M. Plaza, I. L. Markov, and J. P. Hayes, Enhancing design robustness with reliability-aware resynthesis and logic simulation // *Proc. 2007 IEEE/ACM Int. Conf. Comput.-Aided Des.*, 2007, pp. 149–154.
- [14] Stempkovskiy, A., Telpukhov, D., Nadolenko, V. Development of Resynthesis Flow for Improving Logical Masking Features of Combinational Circuits // Proceedings of 2018 IEEE East-West Design and Test Symposium, EWDTS 2018
- [15] A. Glebov, S. Gavrilov, D. Blaauw, V. Zolotov, R. Panda, C. Oh, "False-Noise Analysis using Resolution Method", *ISQED 2002*, March 2002.
- [16] Stempkovskiy A.L., Tel'pukhov D.V., Solov'ev R.A., Myachikov M.V. Metody povysheniya proizvoditel'nosti vychisleniy pri raschete metrik nadezhnosti kombinacionnykh logicheskikh skhem // *Vychislitel'nye tekhnologii*. 2016. T. 21. № 6. (in Russian).
- [17] Mischenko, A., Brayton, R. SAT-based complete don't care computation for network optimization // *DATE '05*. P. 412-417.
- [18] Plaza, S., Chang, K-H., Markov, I., Bertacco, V. Node mergers in the presence of Don't Cares // *ASP-DAC '07*. P. 414-419.

Neural Net as Pseudo-Inverse Filter in Speech Coding Problem

Rustam Latypov
Department of System Analysis and IT
Kazan Federal University
Kazan, Russia
Roustam.Latypov@kpfu.ru

Evgeni Stolov
Department of System Analysis and IT
Kazan Federal University
Kazan, Russia
ystolov@kpfu.ru

Abstract— In the general case, a finite impulse response (FIR) filter has no inverse filter. To solve the inverse filtering problem, we propose an approximate method that restores the initial sequence. We analyze the sequence obtained by filtering the source sequence with an arbitrary FIR filter. The analysis made with the help of a deep neural network. The results of the study are applied to speech coding. We show experimentally, that our approach provides lowering bit rate in the transmission of speech data through a channel.

Keywords— FR filter, pseudo-inverse filter, neural net, speech coding

I. INTRODUCTION

According to the theory, the smallest bit rate, at which the original signal can be transmitted without distortion, is determined by the message entropy. Yet, in practice, the source speed corresponding to the entropy is reachable only asymptotically. Audio compression is related to the efficient handling of audio data with good perceptual quality. Practical methods of source compression use lossy coding, which usually guarantees bit rate savings due to almost imperceptible signal degradation [1].

Audio files require a great deal of bandwidth for processing. Although the bandwidth of digital channels increases every year, the problem of signal compression during transmission stay around actual [2]. Many pieces of research are initiated by this issue. For example, the book [3] gives an extensive range of various speech compression methods, just as the book [4] provides a wide discussion of different signal processing algorithms that are designed to perform noise reduction in speech files. Standards for speech processing are also developed, such as G.711, G.726, G.728, and others [5]. All those standards suppose online processing the signal. It should be noted that all quality losses of the transmitted signal must be perceptually evaluated and the intelligibility is the most important criteria.

If the distortion is recognized, it can be compensated. Restoration is based on a distorted version of the original signal mixed with noise. Inverse filtering as a method of recovering a distorted signal is widely used in image processing, speech processing technologies, and other applications [6 – 8]. Inverse filtering is known to be sensitive to additive noise. For example, a low-pass filter may have zeros or small values at selected

frequencies, so these frequencies will be amplified in noise. The problem is usually solved by approximation inverse filter with pseudo-inverse filter [9–11].

In our paper, we propose a pseudo-inverse filtering method implemented as a deep neural net. An advantage of our method is as follows. The most modern methods of data compression are based on the idea of the codebook. Following the general approach, the source file is divided into segments, then a special algorithm collects the features of the segment, and these functions are sent to the recipient. Having a codebook at endpoint, the receiver looks for the segment that fits the taken features in the codebook and produces a new segment that will be sent to the client. Some new applications of that idea can be found in [12]. When implementing such an algorithm, an additional problem arises, how to glue a sequence of fragments and convert it into one signal. Our method allows us to avoid the bonding procedure.

The paper is organized as follows. First, we provide an introduction. Some definitions and ideas about pseudo-inverse filtering are in the next section. Then we describe the experimental results, give some discussions, and make a conclusion.

II. NEURAL NET AS A PSEUDO-INVERSE FILTER

FIR filters constitute a class of digital filters which are based on a feed-forward difference equation. FIR filters operate on discrete-time signals and the filter output is computed as a weighted, finite sum

$$y_k = \sum_{i=0}^{M-1} x_{k-i} \cdot a_i, \quad k = 0, 1, \dots, N - 1. \quad (1)$$

Here we have designated the real sequence that should be filtered out as x_0, x_1, \dots, x_{N-1} , the numbers a_0, a_1, \dots, a_{M-1} are the FIR filter real coefficients, $M - 1$ is the filter length. As usual, $x_i = 0$, if $i < 0$ in (1).

An inverse filter is implemented to restore the original sequence:

$$x_k = (y_k - \sum_{i=0}^{M-1} x_{k-i} \cdot a_i) / a_0, \quad k = 0, 1, \dots, N - 1. \quad (2)$$

It is known that the difference equation (2) works as a stable filter with an infinite impulse response if the roots of the polynomial

$$f(z) = \sum_{i=0}^{M-1} a_i \cdot z^{M-1-i}$$

lie inside the unit circle [13]. In general, this condition is not satisfied for arbitrary FIR filters, so the inverse filter does not always exist.

Our task is to develop a technique for the approximate restoration of the original sequence in the case of trouble with the design of the inverse filter. In this article, we use a deep neural network to predict that audio sequence using own dataset of audios. We feed the filtered sequence to the deep neural network inputs. Our goal is to show that under suchlike conditions a neural network can be used as a regression function and such a network can be trained to approximate the original sequence.

We describe some heuristic considerations when training and testing the network.

While training, the net has access to the system of equation

$$\begin{cases} a_0x_k + a_1x_{k-1} + \dots + a_{M-1}x_{k-M+1} = x_k \\ a_0x_{k+1} + a_1x_k + \dots + a_{M-1}x_{k-M+2} = x_{k+1} \\ \vdots \\ a_0x_{k+p} + a_1x_{k+p-1} + \dots + a_{M-1}x_{k+p-M+1} = x_{k+p} \end{cases},$$

where the coefficients of the filter are unknown variables. When solving such a system, we get a set of numbers a_0, a_1, \dots, a_{M-1} , so the filter coefficients can be evaluated during training. The system may have no exact solution, but in any case, the net can obtain an approximation of the solution.

After evaluations of the coefficients a_0, a_1, \dots, a_{M-1} , a new system of equations arises with unknown values of x_k :

$$\begin{cases} a_0x_k + a_1x_{k-1} + \dots + a_{M-1}x_{k-M+1} = y_k \\ a_0x_{k+1} + a_1x_k + \dots + a_{M-1}x_{k-M+2} = y_{k+1} \\ a_0x_{k+2} + a_1x_{k+1} + \dots + a_{M-1}x_{k-M+3} = y_{k+2} \\ \vdots \end{cases}.$$

Solving this system, we obtain the source sequence.

In our experiment, we leveraged audio files written with the same sampling frequency as source sequences. The most important observation from the experiments is as follows: the described procedure weakly depends on the source audio sequence x_0, x_1, \dots, x_{N-1} . The trained net can reconstruct the arbitrary audio sequence s_0, s_1, \dots, s_{Q-1} of some length Q when the result of filtering the sequence by the same filter is given. The remarks presented above can explain this phenomenon but these considerations cannot be viewed as a mathematical proof. According to the practice accepted in the theory of neural net, all hypotheses should be tested by experiments. Since the presented results somewhat contradict intuition, we give all the parameters of the functions used in our experiments.

III. EXPERIMENT

A. Make a deep neural network

We created a simple neural net using *Keras* package with *TensorFlow* as backend [14]. The parameters of the net are as follows.

1. Define three layers having 64 neurons each.

2. The input layer has 101 inputs.
3. Define the output layer with a single neuron.
4. Use *ReLU* as the activation function for the hidden layers.
5. Use *MSE* (mean-square-error) metrics to estimate the linear regression.

B. Dataset

All speech arrays are sampled with the frequency 44100 Hz. Denote by *Source* a speech array containing samples of 16 bits. The element having the index N in the target array is *Source*[N]. For filtering *Source*, we use symmetric FIR filter produced by function *FIRWIN* from the package *SciPy* [15]. The filter has length 101, but the result of the approximation significantly depends on the type of the filter. When creating a pseudo-inverse filter, we only change the file for training, as well as the cutoff frequency of the FIR filter, so we can tag any filter with the file name and this frequency.

Let *FSource* be the result of filtering *Source*. The training data consist of records of the length 101. The record with the index N equals the chunk *FSource*[$N : N+101$].

C. Train the network

We train the net using the training data and the target array. We have to emphasize that the net does not know the coefficients of the filter.

After completing the training, we load another speech array called *SourceNew*. We get the array *FSourceNew* which is the result of filtering *SourceNew* with the same filter. Let *InData* be a list of records of the form *FSourceNew*[$N : N+101$]. We apply *InData* to the inputs of the net and played back the predicted data.

D. Some features of the inverse filter

For the inverse filter, let us set a formal feature, which is an analog of the transfer function. We leveraged two files *m.wav* and *f.wav* containing records of male and female voices respectively. Since we are going to use low-pass filters in our application, we need to investigate an analog of the transfer function for such filters although in our case we are dealing with a non-linear filter. Here we introduce the results for a low-pass FIR filter with a cutoff frequency of 120 Hz. The net was trained on the files *m.wav* and *f.wav*, so two pseudo-inverse filters are actually under consideration. We denote them as *FiltM* and *FiltF* respectively. A sinusoidal signal was applied to the network inputs, and the absolute values of the Fourier Transform of the received signals are shown in the next figures (Fig.1 and Fig.2).

For comparison, we present the result on a high-frequency FIR filter with a cutoff frequency of 120 Hz and various input frequencies (Fig. 3 and Fig. 4).

E. Quality of the recovered file

The quality of the restored speech file was evaluated by a human being. It is necessary to note that usage of signal with a low frequency of sampling in range of 16kHz leads to poor quality of the restored signal. To evaluate the quality of the procedure, we used the traditional method based on a signal-to-

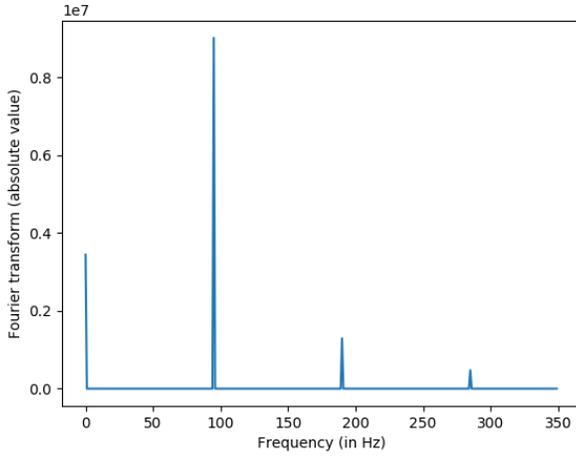


Fig. 1. *FiltM*: response result for a sinusoidal input signal of frequency 95 Hz and low-pass filter with cutoff frequency of 120 Hz.

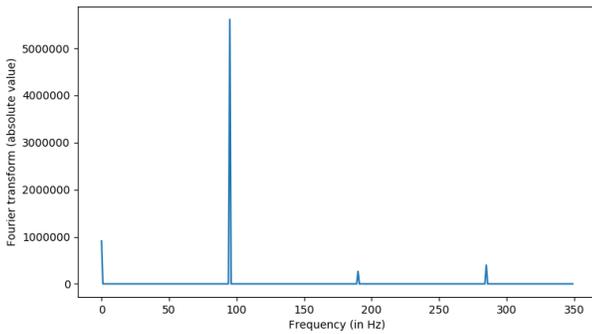


Fig. 2. *FiltF*: response result for a sinusoidal input signal of frequency 95 Hz and low-pass filter with cutoff frequency of 120 Hz.

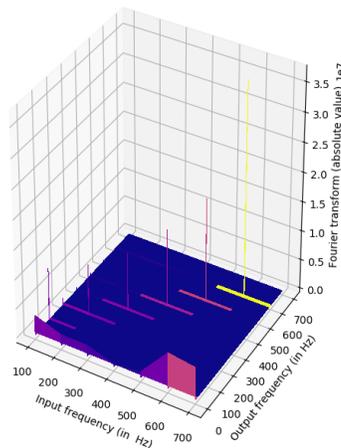


Fig. 3. *FiltM* (3D): experiment for different input frequencies.

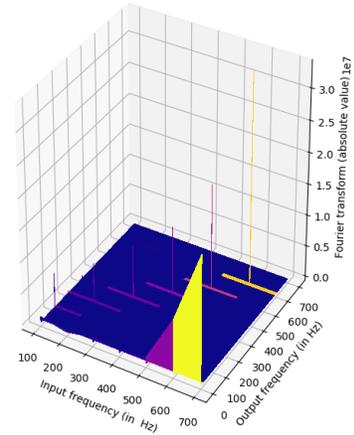


Fig. 4. *FiltF* (3D): experiment for different input frequencies.

noise ratio (SNR). SNR compares a level of signal power to a level of noise power. Our version of SNR function to compare the original file F_1 and the distorted one F_2 after aligning standard deviations has the form

$$SNR = 10 \cdot \log_{10} \left(\frac{\sigma^2(|fft(F_1)|)}{\sigma^2(|fft(F_2)| - |fft(F_0)|)} \right). \quad (3)$$

Here for signal S we use $fft(S) = (S_0, S_1, \dots, S_{P-1})$ – the vector containing results of Fast Fourier Transform of S , and $|fft(S)| = (|S_0|, |S_1|, \dots, |S_{P-1}|)$.

We utilized two files to obtain SNR. We trained the net using the first file and tested the quality of the recovery using the second one (Table I), then we swapped the files in the procedure (Table II). In the tables below the symbol * denotes the file used for training.

TABLE I. SNR. THE NET WAS TRAINED BY M.WAV FILES.

Cutoff frequency	Files to obtain SNR	
	<i>m*.wav</i>	<i>f.wav</i>
75	7.0	4.2
90	7.0	4.2
100	7.4	4.6
110	7.3	4.7
120	7.4	4.6
130	7.5	4.8

TABLE II. SNR. THE NET WAS TRAINED BY F.WAV FILES.

Cutoff frequency	Files to obtain SNR	
	<i>m*.wav</i>	<i>f.wav</i>
75	3.0	8.7
90	3.0	8.8
100	3.4	9.2
110	3.4	9.3
120	3.3	9.2
130	3.5	9.4

The presented results are expectable. The network recovers the signal, part of which was used in neural network training, better than any other signal. On the other hand, the audibility of the recovered signals is acceptable in both cases. Increasing the value of SNR with expanding filter bandwidth is also very natural.

IV. LOWERING BIT RATE

Let us demonstrate the way the presented results can be leveraged to lower bit rate while the speech signal is transmitted through the channel.

First of all, recall the way the lowering of the number of bits is realized in modern protocols [3,12]. The source signal is downsampled with step M . According to Kotelnikov-Shannon theorem, only low frequencies of the signal spectrum are saved as a result. To restore the lost part of the spectrum, some additional information follows the transmitted fragment. The idea of our approach is presented in Fig.5. On the upper side is shown the standard downsampling procedure. On the bottom side, we use upsampling to keep the previous length of the signal.

The next step is setting values at added points. To do that, we use linear interpolation instead of filtering [12]. In place of additional information concerning transmitted fragment, we leverage the neural net that was trained with the filter on the upper side of the channel. To apply a signal to inputs of the neural net one has to convert a signal into a sequence of records of the same format which was used in training the net. The net converts those new records into output signals. We made experiments with the values of M in the interval [4,8] and gained the admissible quality of the signal. The SNR evaluations for cutoff frequency equals 120 Hz are placed in Table III and Table IV.

Comparing Tables I, II, III, IV, one can see the expected results: the SNR estimate in Table I is better than in Table III, and in Table II is better than in Table IV. Although the filtered files have spectra in the interval $[0, 120\text{Hz}]$ and the downsampled files have sample frequency more than the

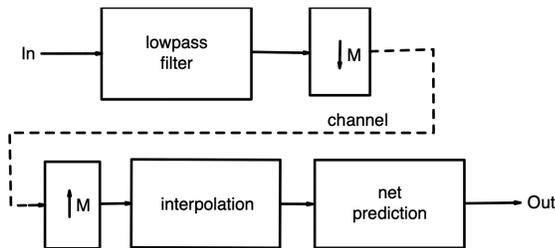


Fig. 5. Lowering bitrate by downsampling and prediction.

TABLE III. SNR AFTER TRANSMISSION, CUTOFF=120 HZ. THE NET WAS TRAINED BY M.WAV FILES.

M	Files to obtain SNR	
	<i>m*.wav</i>	<i>f*.wav</i>
4	4.4	3.9
5	5.2	3.8
6	3.8	3.4
7	2.3	3.3
8	3.2	3.2

TABLE IV. SNR AFTER TRANSMISSION, CUTOFF=120 HZ. THE NET WAS TRAINED BY F.WAV FILES.

M	Files to obtain SNR	
	<i>m*.wav</i>	<i>f*.wav</i>
4	2.3	8.4
5	2.3	8.2
6	2.1	8.2
7	1.6	8.1
8	2.1	7.9

Nyquist frequency, a degree of degradation of SNR can be viewed. The SNR lowers with rising of M . A deviation from this rule, which can be observed in Tables III and IV, can be a result of some features of *m.wav* file.

V. CONCLUSION

The parameters of the neural net, filter, and M that are used in our experiments are far from optimal. We hope that active experiments will bring better results. The goal of this paper is to draw the attention of researchers to the new approach to the coding of audio signals.

REFERENCES

- [1] K. Sayood, Introduction to Data Compression, 5th ed. Elsevier, 2017.
- [2] P. Nizampatnam and K. Kumar, "Bandwidth Extension of Speech Signals: A Comprehensive Review," International Journal of Intelligent Systems Technologies and Applications, vol.2, issue 2, pp.45-52, 2016.
- [3] R. Martin, U. Heute, and C. Antweiler, Advances in digital speech transmission, Wiley Publishing, 2008.
- [4] P. Loizou, Speech Enhancement: Theory and Practice, 2nd ed. CRC Press, 2017.
- [5] J. Gibson, "Challenges in speech coding research," in Speech and audio processing for coding, enhancement and recognition, T. Ogunfunmi, R. Togneri, and M. Narasimha, Eds. Springer, 2015, pp. 19–40.
- [6] K.Wu, F. Ahmed, G. Georgiou, and S. Roumeliotis, "A square root inverse filter for efficient vision-aided inertial navigation on mobile devices," in Proceedings of Robotics: Science and Systems XI (RSS). 2015, MIT Press.
- [7] K. Senmoto and D. Childers, "Signal resolution via digital inverse filtering," IEEE Transactions on Aerospace and Electronic Systems, vol.8, pp.633-640, 1972.
- [8] A. Riaud, M. Baudoin, J. Thomas, and O. BouMatar, "SAW synthesis with IDTs array and the inverse filter: toward a versatile saw toolbox for microfluidic and biological applications," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol.63, p.1601, 2016.
- [9] M. Yeates, "A neural network for computing the pseudo-inverse of a matrix and application to Kalman filtering," Tech. Report, California Institute of Technology, 1991.
- [10] M. Unser and M. Eden, "FIR approximations of inverse filters and perfect reconstruction filter banks," Signal Processing, vol.36, issue 2, pp.163–174, 1994.
- [11] H. Rao, V. Mathews, and Y. Park, "Inverse filter design using minimax approximation techniques for 3-D audio," Proceedings of the IEEE International Conference on Acoustic Speech and Signal Processing, vol. 5, pp. 14–19, 2006.
- [12] A. Oppenheim and R. Shafer, Discrete-time signal processing. Prentice Hill, 1980.
- [13] V. Ramasubramanian and H. Doddala, Ultra low bit rate speech coding. Springer, 2015.
- [14] Keras Homepage, //https://keras.io/. Last accessed 5 April 2019.
- [15] SciPy Homepage, //https://www.scipy.org. Last accessed 5 April 2019.

Protograph Sieving Method for Construction Moderate Length Multi-Edge Type QC-LDPC codes

Svistunov German
 Omsk State Technical University
 Department of Computer Engineering, Omsk, Russia
 Email: g.v.svistunov@gmail.com

Usatyuk Vasilii, Egorov Sergey
 South-West State University
 Department of Computer Science, Kursk, Russia
 Email: L@Lcrypto.com, sie58@mail.ru

Abstract—We introduce the protograph sieving method as a tool for optimization of moderate length Multi-Edge Type Quasi-Cyclic Low-Density Parity-Check (MET QC-LDPC) codes. The proposed method allows to improve code distance property with defined graph balanced cycles and construct MET-LDPC protograph for the code length in gap between Scattering PEXIT-chart and Covariance Evolution methods. We show that this approach improves code distance properties and achieves 0.2 dB gain in terms of block error rate (BLER) over the additive white Gaussian noise (AWGN) channel for information length of 600 bits and rate 1/3, at a target BLER of 10^{-6} when compared to BG2 MET-LDPC code proposed at 5G standard.

Index Terms—Protograph; Protograph Exit-Chart; Covariance Evolution; balanced cycles; Trapping Sets; Sieving; Scattering Exit-Chart; Multi-Edge Type LDPC; MET-LDPC

I. INTRODUCTION

MET QC-LDPC codes applied to several modern standards: 5G eMBB [1], TV physical layer standard ATSC 3.0 [2], Deep Space Communication [3], fiber optic communication [4]. Multi-edge Type LDPC approach is based on the idea of code graph puncturing with special distribution of erasure recoverability [5]. Codes based on MET-approach require more iterations for decoder convergence but perform better iterative decoding threshold (performance as well) [6], [7]. Some construction methods of Multi-edge Type QC-LDPC codes were proposed recently [8]–[14], [21].

Most of these methods have main objective to predict iterative decoding threshold for an asymptotic code length without quasi-cyclic restriction of the Trapping set (cycles and cycles overlap) and code distance. Covariance evolution (Finite-Length scale) predicts well under the negligible influence of cycles overlap, which usually strongly depends upon the code rate and the size of base matrix (circulant size). Finite-Length scale usually works for code rate $rate < 0.9$ and information length about 200–600, [11], [13]. Scattering protograph exits-chart method is suitable for short length $K < 200$ because it requires not so many different degree distributions inside the protograph and allows reasonable time for Monte-Carlo simulation. Main contribution of the paper is an approach for sieving protographs of Short-Length MET QC LDPC codes with improved graph connection properties. Proposed method allows to construct Multi-edge Type protograph for code length in gap between Scattering PEXIT-chart and Covariance

Evolution methods. We will show that this approach achieves coding gain in terms of BLER of 0.2 dB over the AWGN-channel for the codeword length of 1800 bits and a target BLER of 10^{-6} when compared to BG2 proposed at 5G.

II. MULTI-EDGE TYPE QC-LDPC CODES

A usual way to represent QC-LDPC code is to describe by a parity-check matrix H which is made of square blocks. Those blocks could be either zero matrix or circulant permutation matrices. Let the $mL \times nL$ matrix H be defined as follows

$$H = \begin{bmatrix} P^{a_{11}} & P^{a_{12}} & \dots & P^{a_{1n}} \\ P^{a_{21}} & P^{a_{22}} & \dots & P^{a_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ P^{a_{m1}} & P^{a_{m2}} & \dots & P^{a_{mn}} \end{bmatrix},$$

where P^k is the circulant permutation matrix (CPM). CPM is an identity matrix I shifted to the right by k times for any k , $0 \leq k \leq L - 1$. In our notation we define the zero matrix as P^∞ . A set $\{\infty, 0, 1, \dots, L - 1\}$ will be denoted as A_L , $a_{i,j} \in A_L$. Let us denote the circulant size of H as L and consider a code C with parity-check matrix H as a QC-LDPC code.

We call $E(H) = (E_{ij}(H))$ the *exponent matrix* of H defined as:

$$E(H) = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix},$$

i.e., the entry $E_{ij}(H) = a_{ij}$.

Replacing ∞ and other in $E(H)$ by 0 and 1, respectively we obtain the *protograph mother matrix or base graph* $M(H)$ as a $m \times n$ binary matrix.

Let us define a *block-cycle* of length $2l$ as a cycle of length $2l$ in the Tanner graph of $M(H)$. And we define as an *exponent chain* any block-cycle of length $2l$ in $M(H)$ which is both correspond to the sequence of $2l$ CPM's $\{P^{a_1}, P^{a_2}, \dots, P^{a_{2l}}\}$ in H and the sequence of $2l$ integers $\{a_1, a_2, \dots, a_{2l}\}$ in $E(H)$.

There is an easy way to find cycles in the parity-check matrix H Tanner graph. An exponent chain forms a cycle in the Tanner graph of H if the following condition holds, [16]

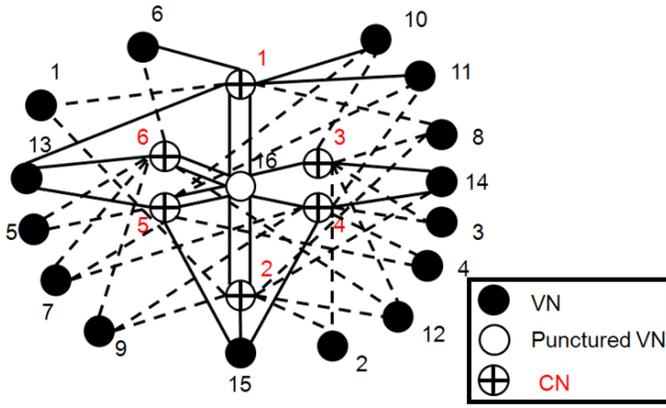


Fig. 1. Graphical representation of MET-LDPC code protograph, $rate = 2/3$

$$\sum_{k=1}^{2l} (-1)^k a_k \equiv 0 \pmod{L}. \quad (1)$$

In equation (1) each coefficient $a_{i,j}$ for CPM $P^{a_{i,j}}$ is added with plus for each even step and with minus for each odd steps. If number of even steps for CPM $P^{a_{i,j}}$ equal to number of odd steps, then $a_{i,j}$ is eliminated from equation (1).

MET LDPC codes generalize the class of irregular LDPC codes. MET approach allows to make a good estimation of iterative decoding threshold based on protograph structure. The edges of MET-LDPC protograph could be divided into several special types according to the connection structure through a CPMs. Each type of edges distributes different messages every iteration. Punctured variable nodes are readily admitted as another important advantages of MET-LDPC codes because its useful both for improving iterative decoding threshold and for lowering error floors, due to properties of cycles cancellation.

On Fig. 1 example of MET-LDPC code protograph with rate $2/3$ are represented. This protograph performs an iterative decoding threshold 1.32 dB while Shannon limit for this case is 1.059 dB. For example, AR4JA MET-LDPC code performs an iterative decoding threshold 1.414 dB, [3].

Balanced cycle on the parity check matrix H is the cycle where each cell participates same times in even steps and in odd steps [20]. Equation (1) for balanced cycle reduces to trivial equality $0 = 0$, which holds for any coefficients $a_{i,j}$. The girth g of the code is the length of the shortest cycle in its Tanner graph and the balanced girth is the minimal length of the balanced cycle. It is clear that the actual girth cannot exceed the balanced girth of a parity check matrix H for any shifts of circulant permutation matrices. That means that balanced girth is an important upper bound for the girth of the protograph. For QC-LDPC codes the bound of balanced cycles was proved at papers [16], [19], [20].

A subgraph in the Tanner graph of H formed by cycles or their overlap with a variable nodes and b odd degree checks is a trapping set (a, b) . For example the overlapping of three 8-cycles gives us the trapping sets TS(5,3) and a TS(4,4) could be formed by cycle 8 in Tanner graph, [17].

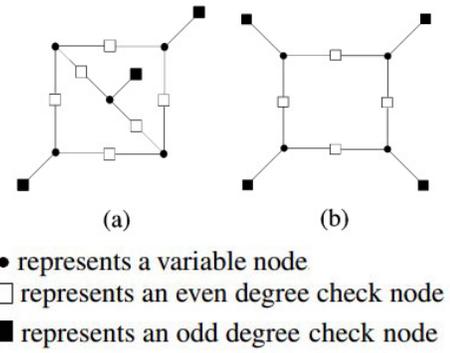


Fig. 2. Graphical representation of Trapping sets: a) TS(5,3), b) TS(4,4)

The number of variable nodes which can cause the fail of the sub-optimal iterative decoding on this TS if corrupted defines its harmfulness. The (4,4) trapping set produces 4 errors in variable nodes after 4 errors in odd degree check nodes. It can be considered as a weight 4 pseudo-codeword. The pseudo-codewords weight spectrum could be improved and the probability of error-floor could be decreased if we will break the most harmful cycles. Some modifications could change the probability of decoding fail caused by TS pseudo-codewords, [18].

In [15] the number of check nodes singly connected to the variable nodes involved in the cycle is defined as Extrinsic Message Degree (EMD) metric of the cycle. As soon as each cycle in the Tanner graph could be considered as a trapping set the EMD metric becomes an important characteristic of a code. It estimates how strong the subgraph of cycle is connected with the rest of the Tanner graph.

Estimation of an block error probability for LDPC code with a change in the code length obtained by the method of the Covariance Evolution (Finite-length scale, [11]) :

$$P_{FER}(N, \sigma) \cong Q\left(\alpha\sqrt{N}(\sigma^* - \sigma)\right), \quad (2)$$

where N is the code length, α is the protograph's ensemble reciprocal scale factor, which approximate number of degree-one check nodes scale under parabolic drift Brownian motion for equivalent to belief propagation peeling decoder, σ^* is the threshold of the iterative decoding (from Density Evolution), σ is the standard deviation in the AWGN-channel, Q - Q function.

The equation is valid for ensembles of codes at lengths for which trapping sets are distributed according to Poisson's law. However, for short lengths MET QC-LDPC codes based on protographs with a σ^* iterative decoding threshold closed to Shannon limit, Poisson's law assumption is not satisfied. Small size of CPM does not allow to get high value of EMD metric.

III. SIEVING METHOD FOR PROTOGRAPH CONSTRUCTION OF MODERATE LENGTH MET QC-LDPC CODES

In proposed sieving MET-LDPC protograph construction method we are searching for some trade-off between graph

and code properties. To solve this optimization problem we propose to generate a protograph according to defined degree distribution or by extending a preliminary optimized core code using greedy methods. By this method a family of protograph candidates could be obtained and a multistage sieving procedure according to the graph properties (balanced cycles) could be executed. Algorithm for balanced cycles search enumerates all closed cycles with length shorter than given K going through non-empty protograph cells with vertical lines on even steps and horizontal lines on odd steps. For each cell of the cycle a difference between appearance after even and after odd step is calculating. Multiplicity of this balanced cycles is defined by protograph automorphism (or the size L of circulant permutation matrix). On practice usually, it is enough to reach girth 8 of balanced cycles to realize code distance which allows to reach $BLER < 10^{-5}$. Another important property is an influence of the protograph and a lifted code to the code distance. Code distance is especially important for error-floor region performance under short and moderate lengths.

At paper [22] there was the proof of the theorem about minimal distance of code defined by parity-check matrix constructed from permutation matrix of arbitrary weight and overlapped permutation matrix.

Theorem 2, [22]. If a parity-check matrix of height M contains a submatrix of height M and width $(j+1)M/j$ contains $j(j+1)$ non-overlapping permutation matrices that all commute each other, then the corresponding code has minimal distance less than or equal to $(j+1)!$. Upper bound of code distance for protograph and tighter bound for lifted codes was generalized and proved by Vontobel and Smarandache, [23].

Permanent of matrix B with size $m \times m$ defined by equation

$$perm(B) = \sum_{\sigma} \prod_{j \in [m]} b_{j, \sigma(j)}.$$

where σ takes all $m!$ permutation of set $[m]$.

Theorem 7, [23]. Let C be QC code defined by the polynomial parity-check matrix $H(x)$. Then the minimum Hamming distance of C is upper bounded as follows

$$d_{min}(C(H)) \leq \min_{S \subseteq [L], |S|=J+1} \sum_{i \in S} wt(perm(H_{S \setminus i}(x))).$$

At paper [24] Butler and Siegel generalized this bound for class of Quasi-Cyclic Multi-Edge Type LDPC codes.

Theorem 12, [24]. Let C' be a QC code constricted by optionally puncturing sub-blocks of the QC code C , defined by the polynomial parity-check matrix $H(x) \in ((F_2[x]/\langle x^N - 1 \rangle)^{J \times L})$ and let $\triangleq wt(H(x))$. Let the sub-block of C indexed by the set P , $P \subset [L]$, be punctured, while maintaining the dimension of the code. Let A' be a submatrix of A with rows a_t , $t \in \tau \subset [J]$, removed. Let S be a subset of $[L]$ of size $J+1 - |\tau|$, such that the subrows $a_{t,S} = 0 \forall t \in \tau$. Then

$$d_{min}(C') \leq \min_{S, \tau} \sum_{i \in S \setminus P} perm(A'_{S \setminus i}).$$

We provide parallel implementation of code distance estimation method for both fixed and variable circulant sizes. A

simultaneous analyze of code properties (code distance) and graph properties (girth, EMD) also is provided [26].

For protograph sieving method as input we use either a core matrix optimized according to the hardware restrictions or a matrix randomly/greedy generated from the degree distribution. Density evolution defines an optimized protograph degree distribution. According to this distribution an ensemble of protographs could be build by different fillings of rows and columns. Real properties of balanced cycles and upper bound of the code distance should be obtained for every protograph from the constructed ensemble. To choose a protograph from the ensemble of candidates we use Protograph-EXIT chart implementation, [26]. Proposed sieving algorithm is defined by the following pseudo code, Alg.1.

Algorithm 1 Sieving method for MET QC-LDPC code protograph construction

Require: Set of Protograph candidate $\{Proto\}$, P_{error} - required level of error probability, Eb/No - required signal to noise ratio to achieve P_{error}

- 1: Estimate upper bound on code distance d_{min}^{upper} for each Protograph using [23], [24], use CPM size as multiplicities.
- 2: **if** $P_{UB} + \delta > P_{error}$ for EB/No **then**
- 3: finish algorithm and generate more Protograph candidate or change degree distribution/size of protograph
- 4: **else**
- 5: Enumerate balanced cycles using test set lifting, weighed TS and estimate union bound for both codeword and trapping sets $P_{UB}^{d_{min}+TS}$
- 6: **end if**

return return base matrix which defines protograph of MET-LDPC codes

In Alg. 1 δ -penalty for the suboptimality of the decoder due to sublinear and linear dependence of TS size from the EB/No . P_{UB} -union bound. Union Bound approximation based on the first term of weight spectrum enumerator for AWGN-channel is given as

$$P_{UB} \approx mult_{d_{min}} Q(\sqrt{d_{min}}),$$

where $mult_{d_{min}}$ - multiplicities of the low weight codewords (according circulant size), $Q(\cdot)$ - Q function.

Below we describe main sieving parameters. If step 1 take too much time, we apply random permutation and try to estimate d_{min}^{upper} again. We shall search our code among the generalization of Raptor based LDPC codes close related to the solution used in eMBB 5G standard [8]. Using proposed sieving method we have constructed MET QC-LDPC protograph with decoding threshold $SNR = -1.57$ for $BER = 10^{-8}$, Table I. Finally we use simulated annealing method which allows to lift the protograph with girth 8, minimal value of EMD 10 (120 cycles were found) and code distance bound 68, [25]. BG2 MET QC-LDPC code with code distance 30, lifted with girth 6, from BG2 family 15 [1], achieves decoding threshold $SNR = -1.7$ for $BER = 10^{-8}$ demonstrate

Calculating the Parameters of the Short-Range Microwave Information Channel Resistant to Signal Fading

Vladimir Mikhaylovich Artyushenko
Information technology and management systems department
Technological university
Korolev city, Russian Federation
artuschenko@mail.ru

Vladimir Ivanovich Volovach
Informational and electronic service department
Volga region state university of service
Togliatty city, Russian Federation
volovach.vi@mail.ru

Abstract—Using a short-range microwave radio line allows to ensure high noise-resistance during transmitting discrete information to the vehicle with relative ease of technical implementation, including the case when the vehicle is in motion. Energy calculation of the short-range radio line is carried out. Calculation of the noise-resistance of a short range duplex channel resistant to signal fading for coherent detection is carried out. A family of dependences of the signal power at the receiver input on relative positions of the receiving and transmitting antennas is obtained. The dependences of the error probability from the transmitted signal element on the signal-to-noise ratio are found for two practically significant cases: with and without the dominant brilliant point. By using the obtained expressions for both models, the dependences of the error probability on the signal-to-noise ratio are constructed for different parameters of the reflected signal and the directional characteristics of the transmitting and receiving antennas. It is shown that with fading of the microwave signal the error probability in the duplex channel can reach values smaller than $10^{-6} \dots 10^{-8}$ per sign.

Keywords—microwave range, short-range radio line, error probability, coherent reception, horn antenna, signal fading.

I. INTRODUCTION

In the last two decades in the technologically developed countries the issue associated with the use of automated control for various kinds of vehicles, such as buses, taxis, trams, trolleybuses, subway trains, etc. [1, 2, etc.] has been of a vital importance. In the nearest future, intelligent traffic control is implemented and various ground unmanned vehicles, primarily cars, are being used.

One of the most important links in such automated control systems of the traffic is the data exchange channel. Analysis of the issue shows that it is quite effective to transmit current and operational data, and also to adjust the program of vehicle motion in some key points of routes (primarily intermediate stops, stations, etc.) by using the short-range radio line (SRL) of microwave range [3, 4]. In the SRL discrete information is exchanged between the on-board module located on the vehicle and the stationary module. Typically, the probability of an error at reception P_0 should be less than or equal to $P_0 \leq 10^{-6}$ a decimal point. The amount of information transmitted in 10 seconds time can reach tens of Mbit.

The use of microwave information channel has a number of advantages with regards to both noise immunity and technical implementation [5-7, etc.]. Herewith, the distance between the transmitting and receiving devices of the SRL, as a rule, tends to be minimized (several meters). The microwave channel is basically intended to transmit various

telemetric information. Notably, the most convenient moment for the exchange of information is the moment when the vehicle is stationary; however, if necessary, data exchange can be carried out during motion of the vehicle.

The purpose of this work is to carry out energy calculation of the short-range radio line and noise immunity of the short-range duplex channel resistant to signal fading during coherent reception.

Requirements and organization of data transmission system through the SRL, as well as engineering calculation used in the microwave radioline of antennas and the amount of information exchange between the onboard and stationary modules of the system of discrete data transmission were earlier considered in detail by the authors [8-10].

II. ENERGY CALCULATION

Energy calculation is meant to determine the signal power at the receiving point P_r and to calculate the ratio between the signal energy E_s and the spectral density of the noise power N_0 :

$$q^2 = E_c^2 / H_0^2. \quad (1)$$

We calculate the signal power at the input of the receiver with a given output power of the transmitter by the formula [1]

$$P_r = P_{tr} G_{tr} G_r \eta_{tr} \eta_r \frac{\lambda^2}{16\pi^2 R_0^2},$$

where P_{tr} is the power of the transmitter, W; G_{tr} , G_r are power gain of the transmitting and receiving antenna respectively; η_{tr} , η_r are the efficiency of the antenna-feeder path of the transmitting and receiving device respectively; R_0 is the distance between the points of transmission and reception, m.

Fig. 1 shows the dependence (the central graph in the plane $P_r R_0$) for the case when the vehicle stops and the axis of antenna pattern (AP) of the transmitting and receiving antennas are the same, while: $P_{tr} = 5 \times 10^3$ W; $\eta_{tr} = \eta_r \approx 1$; $\lambda = 3 \times 10^{-2}$ m.

The presented dependencies show that with the increase of distance between the receiver and the transmitter the signal power at the receiver input decreases exponentially. Thus, when $R_0 = 10$ m, the signal power is 17.5×10^{-7} W. When $R_0 = 50$ m it decreases down to 0.7×10^{-7} W.

In practice, the place of the vehicle stop is accidental and can be limited only to a certain area $\pm \Delta R = L$. As a result the AP axes of the transmitting and receiving antennas may not coincide. In this case, the signal power at the receiver input can be found, based on the expression

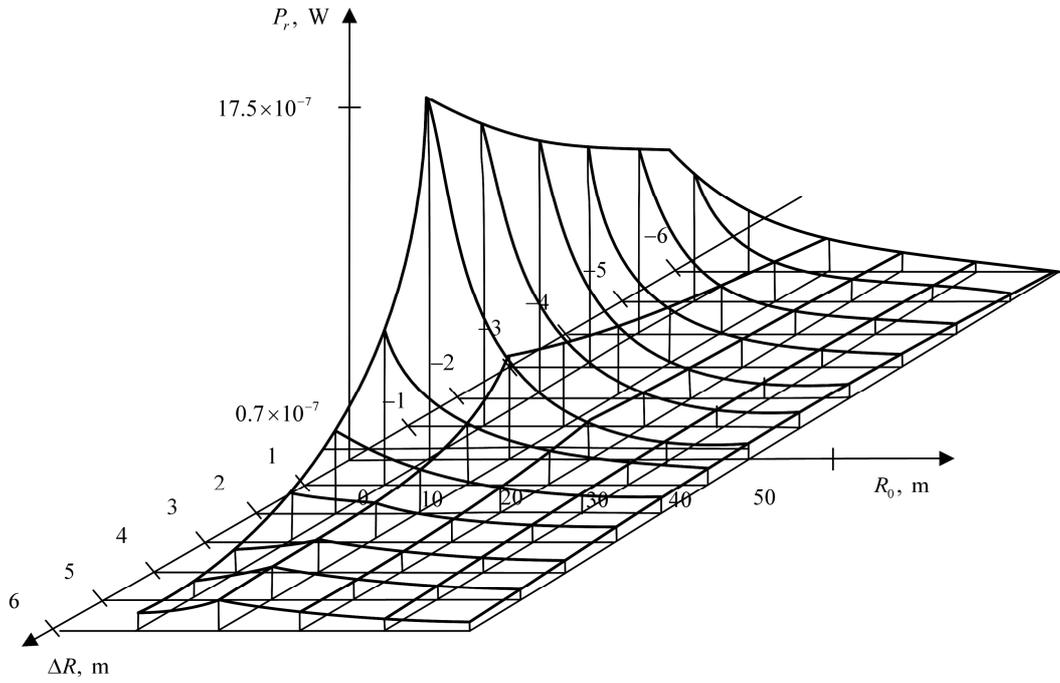


Fig. 1. The dependence of the signal power at the input of the receiver

$$P_r = \frac{K}{R_x^2} F(\alpha) F(\beta),$$

where $K = P_{tr} G_{tr} G_r \eta_{tr} \eta_r \frac{\lambda^2}{16\pi^2}$; $R_x = \sqrt{y_0^2 + (x_0 + x)^2}$; x_0, y_0 are the coordinates of the stationary module antenna characterized by the AP $F(\alpha)$; x is the current location of the vehicle antenna characterized by the AP $F(\beta)$.

Dependencies $P_r = f(R_0; \Delta R)$ on the distances R_0 and ΔR are also shown in Fig. 1 for the case when the axis of the AP of the antenna do not coincide.

III. NOISE IMMUNITY OF THE SHORT-RANGE DUPLEX CHANNEL RESISTANT TO FADING OF THE SIGNAL DURING THE COHERENT RECEPTION

When a duplex channel works at short distances, its noise immunity is significantly affected by fading of signals. It appears as a result of the interference of direct electromagnetic field \dot{E}_r and electromagnetic field \dot{E}_{ref} reflected from the underlying surface in the receiving aperture.

The total field \dot{E}_Σ during a single reception can be described by the ratio

$$\begin{aligned} \dot{E}_\Sigma = \dot{E}_r + \dot{E}_{ref} = 0,5\sqrt{30P_1\eta_1G_1}F_r(\alpha) \times \\ \times \exp\left\{-i\frac{2\pi}{\lambda}r_1\right\}F_r(\beta) + \int_{D(x,y)} |\Phi(x,y)| \frac{\sqrt{30P_1\eta_1G_1}}{r_2} \times \\ \times F_r(\alpha + \Delta\alpha)F_r(\beta + \Delta\beta) \times \\ \times \exp\left\{-i\frac{2\pi}{\lambda}r_2 + i\beta_{ph}\right\} ds(x,y), \end{aligned} \quad (1)$$

where P_1, η_1, G_1 are respectively, the radiation power, the efficiency and the gain of the antenna-feeder path of the direction characteristics of the transmitting and receiving

antenna; α is the angle between the direction of the radiation maximum and the direction to the center of the receiving aperture; β is the angle between the maximum AP of the receiving antenna and the beam towards the transmitting antenna; the angles $\Delta\alpha, \Delta\beta$ and $r_1 = AB$ and $r_2 = AC + CB$ are explained in Fig. 2; $|\Phi(x,y)|$ is the module of the reflection coefficient taking into account the reduction in the amplitude of the wave reflected from the underlying surface; β_{ph} is the phase of the reflection coefficient taking into account phase change during reflection; $C(x,y)$ is the current point of the underlying surface $D(x,y)$ forming the reflected wave.

In amplitude factors it can be considered that $r_1 \approx r_2 \approx r$. Then, the expression (1) can be transformed into:

$$\dot{E}_\Sigma = \dot{E}_0 [F_r(\alpha)F_r(\beta) + J_1(\alpha, \Delta\alpha, \beta, \Delta\beta)] = \dot{E}_0 V,$$

where

$$\begin{aligned} J_1(\alpha, \Delta\alpha, \beta, \Delta\beta) = \sum_{i=1}^N \int_{D(x,y)} F_r(\alpha + \Delta\alpha(x,y)) \times \\ \times F_r(\beta + \Delta\beta(x,y)) \times |\Phi(x,y)| \times \\ \times \exp\{i[k\Delta r(x,y) + \beta_{ph}(x,y)]\} ds(x,y); \end{aligned}$$

$\dot{E}_0 = \frac{1}{r} \sqrt{30P_1\eta_1G_1} \exp\left\{i\frac{2\pi}{\lambda}r_1\right\}$ is a field density during propagation in free space; $k = \frac{2\pi}{\lambda}$; $\Delta r(x,y)$ is the distance up to the current point $C(x,y)$ of the underlying surface $D(x,y)$ forming the reflected wave.

It is shown in [6] that the expression in square brackets is an attenuation factor characterizing the interference of direct and reflected waves.

In practice, the module of attenuation factor is of greatest interest

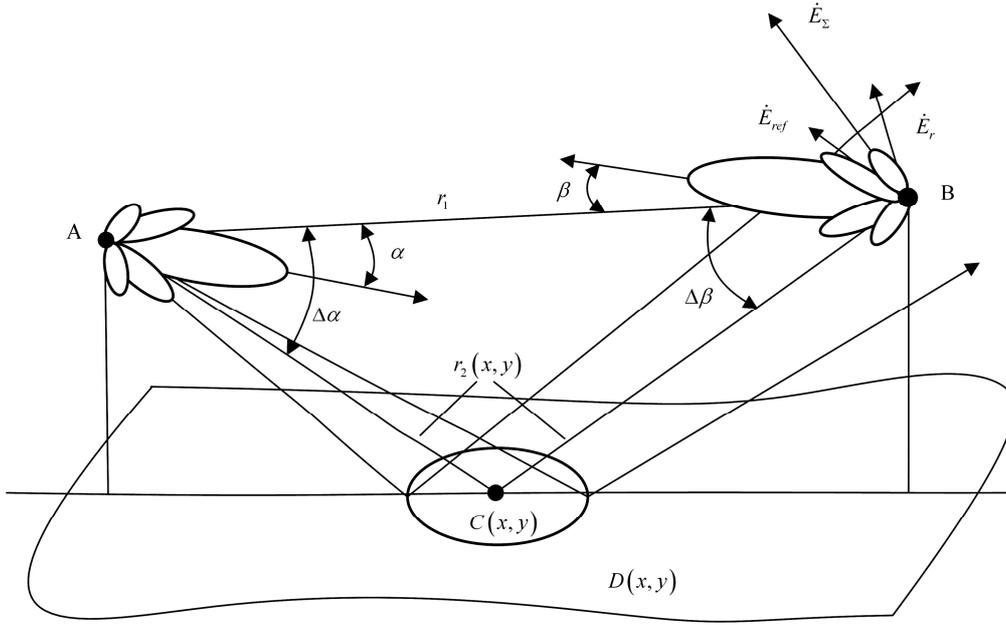


Fig. 2. Process of reflection from the underlying surface

$$|V| = \sqrt{F_{tr}(\alpha)F_r(\beta) + \text{Re}J_1(\alpha, \Delta\alpha, \beta, \Delta\beta)^2 + \text{Im}J_1(\alpha, \Delta\alpha, \beta, \Delta\beta)^2}$$

In order to calculate $J_1(\cdot)$ we use the stationary phase method by presenting $D(x, y)$ as a set of N parts, where the wave can be considered locally flat within each part.

In this case

$$J_1(\alpha, \Delta\alpha, \beta, \Delta\beta) = \sum_{i=1}^N \int_{D(x,y)} F_{tr}(\alpha + \Delta\alpha(x, y)) \times F_r(\beta + \Delta\beta(x, y)) \times |\Phi(x, y)| \exp\{i[k\Delta r(x, y) + \beta_{ph}(x, y)]\} = \sum_{i=1}^N F_{tr}(\alpha + \Delta\alpha(x_i, y_i)) \times F_r(\beta + \Delta\beta(x_i, y_i)) |\Phi(x_i, y_i)| \times \exp\{i[k\Delta r(x_i, y_i) + \beta_{ph}(x_i, y_i)]\},$$

where (x_i, y_i) are coordinates of the stationary phase point of the local part $D_i(x, y)$.

It is quite clear that when the size of the reflected surface $D(x, y)$ is much smaller than the distances r_1 and r_2 , $N = 1$, then the attenuation factor can be transformed into

$$|V| = \left(F^2(\alpha, \beta) + F(\alpha, \Delta\alpha, \beta, \Delta\beta)^2 |\Phi(x, y)| 2F(\alpha, \beta) \right) \times F(\alpha, \Delta\alpha, \beta, \Delta\beta) |\Phi(x, y)| \cos \varphi. \quad (2)$$

Here $F^2(\alpha, \beta) = F_{tr}(\alpha)F_r(\beta)$; $F(\alpha, \Delta\alpha, \beta, \Delta\beta) = F_{tr}(\alpha + \Delta\alpha)F_r(\beta + \Delta\beta)$; φ is the phase of the signal.

Next, we proceed to analyzing the noise immunity in the duplex channel.

The probability of error of the transmitted signal element when using amplitude modulation and incoherent reception is determined by the ratio [7]:

$$P_e = \frac{1}{2} \exp\left\{-\frac{q^2}{2}|V|^2\right\}, \quad (3)$$

where $q^2 = P_s T / N_0^2$; P_s is the power of the element of the signal with duration T ; N_0^2 is a spectral density of normal white noise.

Note that in this case q^2 plays the role of the signal to noise ratio (SNR).

We denote in (2) $A_0 = F^2(\alpha, \beta)$; $B_0 = F^2(\alpha, \Delta\alpha, \beta, \Delta\beta) |\Phi(x, y)|$, then we transform (3) into

$$P_e = \frac{1}{2} \exp\left\{-\frac{q^2}{2}(A_0^2 + B_0^2 + 2A_0B_0 \cos \varphi)\right\}.$$

As a rule, the following assumption is made about the phase distribution

$$\varphi \in W(\varphi) = \begin{cases} \frac{1}{2\pi}, & \varphi \in [-\pi, \pi]; \\ 0, & \varphi \notin [-\pi, \pi], \end{cases}$$

where $W(\varphi)$ is the probability density function (PDF) of the phase.

Taking into account the above assumptions, the probability of error will be determined as:

$$P_e\left(\frac{q}{A_0, B_0}\right) = \int_{-\pi}^{\pi} P_e\left(\frac{q}{A_0, B_0, \varphi}\right) W(\varphi) = \frac{1}{4\pi} \exp\left\{-\frac{q^2}{2}(A_0^2 + B_0^2)\right\} I_0(A_0 B_0 q^2), \quad (4)$$

where $I_0(\cdot)$ is a Bessel function of zero order.

The expression (4) characterizes the random nature of the envelope of the signal reflected from the underlying surface.

Since $D(x, y)$ can be considered as an extended surface, then the same assumptions are true with regards to the PDF $W(B_0)$ as for extended targets.

Two reflection models are of interest:

– with a dominant brilliant point

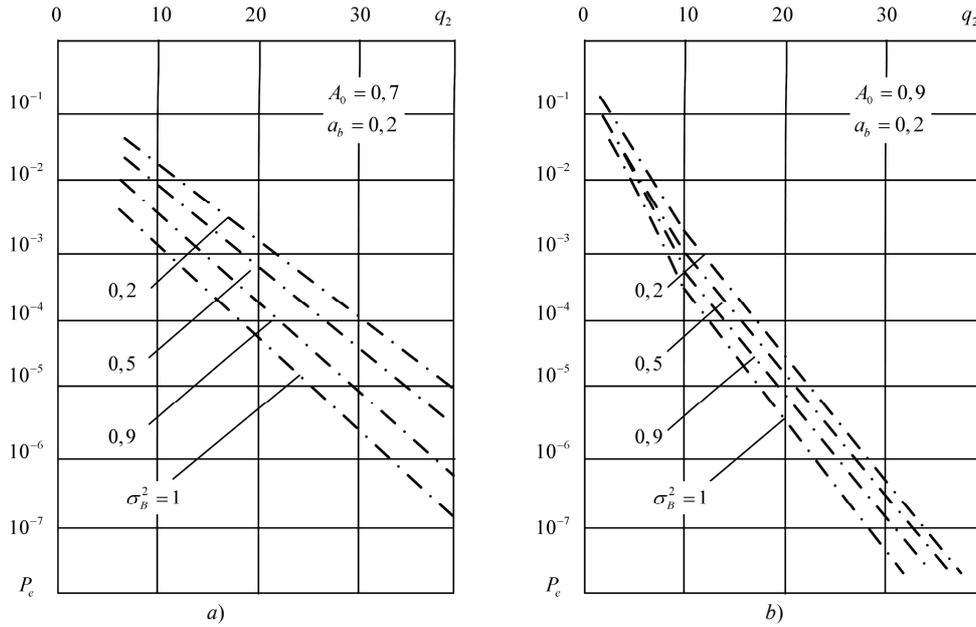


Fig. 3. The dependence of the error probability on the value of the SNR

$$W(B_0) = \frac{B_0}{\sigma_F^2} \exp\left\{-\frac{B_0^2 + a_b^2}{2\sigma_F^2}\right\} I_0(B_0 a_b), \quad (5)$$

where σ_F^2 is the variance of fluctuations of the signal reflected from the surface $D(x, y)$; a_b characterizes the amplitude of the signal reflected from the dominant brilliant point;

– without a dominant brilliant point

$$W(B_0) = \frac{2m^m B_0^{2m-1}}{\Gamma(m) \sigma_F^{2m}} \exp\left\{-\frac{mB_0^2}{\sigma_F^2}\right\}, \quad (6)$$

where m is a distribution parameter; $\Gamma(\cdot)$ is a gamma function.

The ratio (6) is called the Nakagami distribution and it is more general than the Rayleigh distribution. При $m = 1$, the expression (6) transforms into Rayleigh law, whereas when $m > 1$ it can be used even for approximation of the generalized Rayleigh law.

We are going to consider these cases.

Having averaged (4) over the parameter B_0 , which complies to the distribution law (5), we find that the probability of error in this case will be determined as:

$$P_e = \frac{1}{4\pi\sigma_F^2} \exp\left\{-\frac{q^2 A_0^2}{2} - \frac{a_b^2}{2\sigma_F^2}\right\} \times \\ \times \int_0^\infty B_0 \exp\left\{-B_0^2 \left(\frac{q^2}{2} - \frac{1}{2\sigma_F^2}\right)\right\} I_0\left(\frac{B_0 a_b}{\sigma_F}\right) I_0(A_0 B_0 q) dB_0.$$

Taking into account the fact that the integral on the right side is standard [8]

$$\int_0^\infty x \exp\{-\rho x^2\} I_\nu(\alpha, x) I_\nu(\beta, x) dx = \\ = \frac{1}{2\rho^2} \exp\left\{-\frac{\alpha^2 + \beta^2}{4\rho^2}\right\} I_\nu\left(\frac{\alpha\beta}{2\rho}\right).$$

Then for P_e we get

$$P_e = \frac{\exp\{-Q\}}{2\pi\sigma_F^4} I_0\left(\frac{A_0 a_b}{\sigma_F^2 q}\right), \quad (7)$$

where $Q = \frac{(q^4 A_0^2 \sigma_F^4 + q^4 \sigma_F^2 a_b^2 + q^2 2A_0^2 \sigma_F^4 + 2a_b)}{2q^4 \sigma_F^4}$.

When $(a_b^2 + A_0^2) \ll q$, the expression (7) can be simplified $\sigma_F^2 = 1$

$$P_e = \frac{\exp\left\{-\frac{q^2 A_0^2 \sigma_F^2 + a_b^2}{2\sigma_F^2}\right\}}{2\pi\sigma_F^2 q^4} I_0\left(\frac{A_0 a_b}{\sigma_F q}\right).$$

The dependences $P_e = f(q^2)$ with different values A_0 , a_b and σ_F^2 are shown in Fig. 3, with different values of A_0 : $a = 0,7$; $b = 0,9$.

Let us consider the case when $W(B_0)$ is described by the expression (6)

$$P_e = \frac{2m^m \exp\left\{-\frac{q^2 A_0^2}{2}\right\}}{4\pi\Gamma(m)\sigma_F^{2m}} \times \\ \times \int_0^\infty B_0^{2m-1} \exp\{-B_0^2\} I_0(A_0 B_0 q) dB_0,$$

where $\chi = \frac{q^2}{2} + \frac{m}{\sigma_F^2}$.

Using the ratio [8]

$$\int_0^\infty x^\mu \exp\{-\alpha x^2\} I_\nu(\beta, x) dx = \\ = \frac{\beta^\nu \Gamma\left(\frac{1}{2}\nu + \frac{1}{2}\mu + \frac{1}{2}\right)}{2^{\nu+1} 2^{0,5(\mu+\nu+1)} \Gamma(\nu+1)} {}_1F_1\left(\frac{\nu+\mu+1}{2}; \nu+1; -\frac{\beta^2}{4\alpha}\right),$$

where ${}_1F_1(\alpha, \beta, \gamma)$ is a degenerate hyperbolic function.

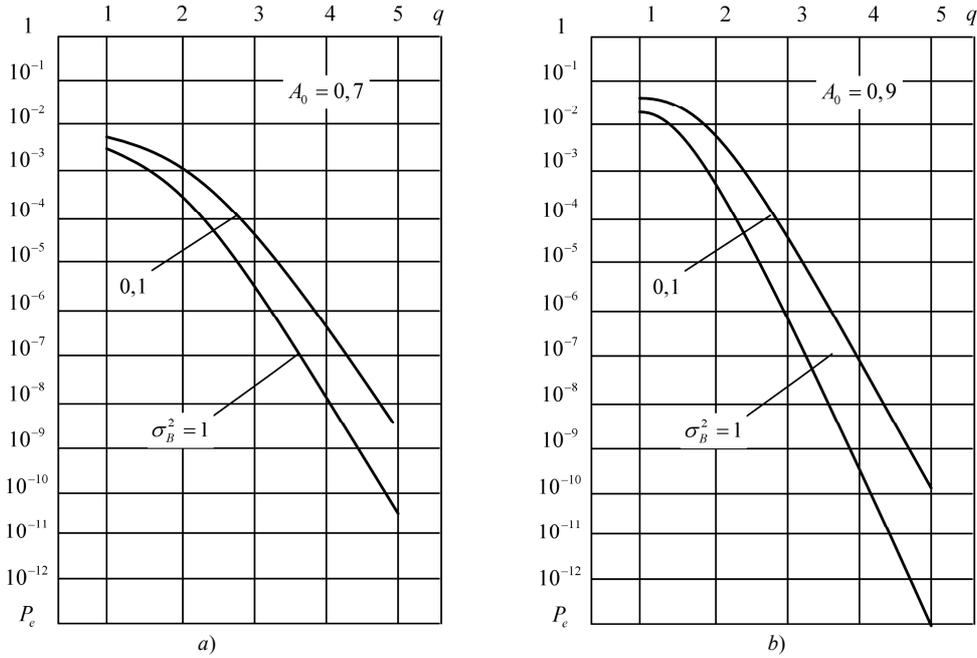


Fig. 4. The dependence of the error probability on the SNR value with different values of σ_B^2

Then we finally write

$$P_e = \frac{m^m \exp\left\{-\frac{q^2 A_0^2}{2}\right\}}{4\pi\sigma_F^{2m}\chi^m} {}_1F_1\left(m, 1, -\frac{A^2 q^4}{4\chi}\right).$$

When $m = 1$, i.e. $W(B_0)$ is a Rayleigh distribution, the following equality is true:

$${}_1F_1\left(1, 1, -\frac{A^2 q^4}{4\chi}\right) = \exp\left\{-\frac{A_0^2 q^4}{4\chi}\right\}.$$

Therefore

$$P_e = \frac{\exp\left\{-\frac{q^2 A_0^2}{2}\left(1 + \frac{q^2}{2\chi}\right)\right\}}{4\pi\sigma_F^2\chi}.$$

The dependences $P_e = f(q)$ for different values of A_0 and σ_F^2 are shown in Fig. 4, for the cases when: $a - A_0 = 0,7$; $b - A_0 = 0,9$.

IV. CONCLUSIONS

The calculation of a number of parameters for the duplex microwave radio line, which serves for data exchange between the vehicle module and the stationary ground station module, is carried out.

Energy calculation of the short-range radio line is carried out. Noise-immunity of the short-range duplex channel resistant to signal fading for coherent detection is analyzed. The results of noise-resistance simulation have shown, that with fading of the microwave signal, the error probability in the duplex channel can reach values smaller than $10^{-6} \dots 10^{-8}$ per sign.

The practical value of the research consists in the obtained dependences of the signal power at the input of the signal receiver for two common cases of the location of the receiving and transmitting antennas allow the calculation of

the information microwave channel. It is also possible to determine the noise immunity of the channel and the probability of error of the transmitted signal element.

REFERENCES

- [1] Van Trees, K. Bell, and Z. Tiany, Detection Estimation and Modulation Theory, 2nd Edition, Part I, Detection, Estimation, and Filtering Theory. London: Wiley & Sons, Inc., 2013.
- [2] V. P. Tuzlukov, Signal Processing Noise, Boca Raton, London, New York, Washington D.C.: CRC Press, Taylor & Francis Group, 2002.
- [3] Ellingson, Radio System Engineering. Cambridge University Press, 2016.
- [4] M. Barkat, Signal Detection and Estimation. Norwood: Artech House, 2005.
- [5] G. L. Charvat Small and Short-Range Radar Systems. CRC Press, 2014.
- [6] Jian Lan, and X. Rong Li, "Tracking of Extended Objects with High-Resolution Doppler Radar," IEEE Trans. on Aerospace and Electronic Systems, 2016, Volume 52, Issue 6, pp. 2973–2989. <https://doi.org/10.1109/TAES.2016.130346>
- [7] Jian Lan, and X. Rong Li, "Tracking of Extended Objects with High-Resolution Doppler Radar," IEEE Trans. on Aerospace and Electronic Systems, 2016, Volume 52, Issue 6, pp. 2973–2989. <https://doi.org/10.1109/TAES.2016.130346>
- [8] V. M. Artyushenko, and V. I. Volovach, "Measuring information signal parameters under additive non-Gaussian correlated noise", Optoelectronics, Instrumentation and Data Processing, 2016, Vol. 59, No. 6, pp. 22-28. DOI: [10.15372/AUT20160603](https://doi.org/10.15372/AUT20160603)
- [9] V. M. Artyushenko, and V. I. Volovach, "Information characteristics signals and noise with non-Gaussian distribution", Proceedings XI International IEEE Scientific and Technical Conference "Dynamics of Systems, Mechanisms and Machines (Dynamics)", Nov. 2017. DOI: [10.1109/Dynamics.2017.8239430](https://doi.org/10.1109/Dynamics.2017.8239430)
- [10] V. M. Artyushenko, V. I. Volovach, and V. N. Budilov, "Synthesis and analysis of discriminators meter information parameters signal under non-Gaussian noise with band-limited spectrum", Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2017). Novi Sad, Serbia, Sept 29-Oct 2, 2017. – Kharkov: KNURE, 2017. P. 355-358. DOI: [10.1109/EWDTS.2017.8110112](https://doi.org/10.1109/EWDTS.2017.8110112)

The Implementation of the Genetic Algorithm Using Cloud-Based Computing on the Internet

Kureichik V. M.

Autonomous Federal State Institution of Higher Education
Southern Federal University
Taganrog, Russia
Vmkureichik@sfedu.ru

Logunova J.A.

Autonomous Federal State Institution of Higher Education
Southern Federal University
Taganrog, Russia
Julia1000@yandex.ru

Abstract— Two important tasks are considered in this paper: highlighting extreme subsets and finding cliques of a graph. These tasks have a great practical importance in the design of automation devices and computing equipment, as well as the creation of control systems, computers and networks, in sociology and game theory. Because they are belong to the class of NP-hard problems, the study, creation and modification of methods for their solution still remains relevant. In connection with the expansion of the information space and the increase information for analysis in times, it becomes obvious that in practice it is advisable to solve such NP-difficult problems using new IT technologies - cloud computing. In this regard, the fundamental difference of this work is not only the construction and verification of the genetic algorithm modification, but also the use of the cloud platform as an Internet service for performing calculations. The main goal of a research is to solve the problem of finding extreme subsets in a graph and building cliques using modified genetic algorithm in a reasonable time. The principal difference of the proposed method is that it considers the population degradation degree and increases the diversity of the population using various genetic operators in the evolutionary adaptation block. Several experiments were also conducted, during which a comparison of the developed algorithms with various crossovers was carried out: ordering crossover and greedy. The practical results of the research almost coincided with the theoretical background.

Keywords—genetic algorithm, extreme subsets, cliques, cloud-based computing

I. INTRODUCTION

In the modern world, due to the increasing competition and the rapid pace of IT technologies development, organizations are increasingly resorting to the cloud technologies using. In essence, this is a fundamentally new concept that allows individual entrepreneurs and organizations to provide and receive IT resources in the form of services.

The cloud infrastructure must meet the following characteristics: all resources are pooled, a self-service at the request of a service consumer, broadband network access, responsiveness and a very important characteristic reflecting the economic benefits of using cloud computing — payment for services for actual consumption. The benefits that an organization receives from cloud computing are obvious: a

decrease in the IT services cost, while the quality of these services tends to increase; business adaptability for market changes, flexible scaling of computing resources, high availability. The services that are provided by cloud systems are measured using special abstract parameters that vary depending on the category and type of service. For example, quantitatively can be estimated computing power, throughput, data storage size [1].

Within the concept of cloud systems, it is customary to distinguish three service models: cloud software as a service (Cloud Software as a Service, SaaS), cloud platform as a service (Cloud Platform as a Service, PaaS) and cloud infrastructure as a service (Cloud Infrastructure as a Service, IaaS). As further we will talk about PaaS, we consider the structure of this model in more detail. It is also the basis for the other two models. PaaS provides the opportunity to place and implement in the cloud infrastructure the application which was already created or purchased by the customer. At the same time, the platform offers the necessary tools for the user, testing tools, database management systems. The user is granted the right to manage this application, as well as the configuration settings of the environment in which the application is deployed. Examples of this model are Google App Engine and Microsoft Windows Azure Platform. The scheme of PaaS is presented on fig. 1 [2].

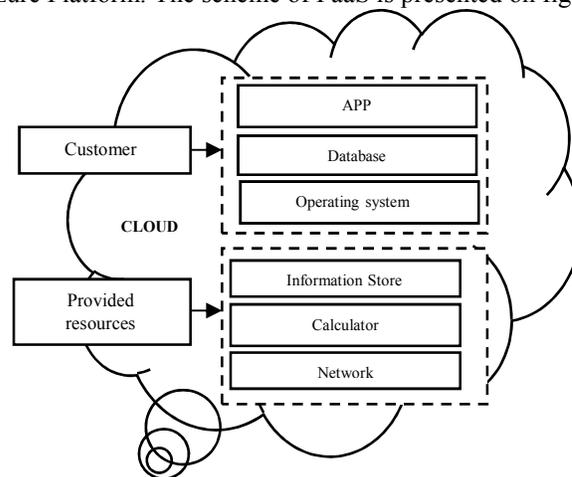


Fig. 1. Cloud Platform as a service model

The cloud systems emergence has significantly improved the availability of services, computing and computer resources, ensure the scalability and flexibility of systems deployed in the cloud environment, and reduce the risks associated with the inoperability and maintenance of infrastructure elements [1]. The task of this work is to solve the problem of isolating extreme subsets and cliques in a graph by a modified genetic algorithm using cloud computing on the Internet. The first task can be useful in the design of automation devices and computing equipment, as well as the creation of control systems [3]. Finding clicks in a graph is used in cluster analysis, particularly in sociology and information retrieval, as well as in game theory [4]. These tasks belong to the NP-difficult problems class and their calculation in a large graph even with heuristic algorithms may require a sufficiently large amount of memory and computational power. In this regard, it may be economically feasible to solve this problem using cloud computing on the Internet.

The idea of using the principles of biological evolution to solve optimization problems arose in the 60s of the last century. And Holland's first book, *Adaptation in Natural and Artificial Systems*, containing the genetic algorithm was published in 1975 [5]. A huge foreign and domestic bibliography shows that genetic algorithms have proved to be an effective method for solving many applied problems [6–8]. In this regard, a genetic algorithm was chosen among the many heuristic search methods to solve the tasks. The main goal of a research is to solve the problem of finding extreme subsets in a graph and building cliques using modified genetic algorithm in a reasonable time.

II. FORMULATION OF PROBLEMS

Let the graph be given $G = (Y, M)$, the subset $T \subset Y$ called intrinsically stable or independent if $T \cap MT = \emptyset$, that is, no two vertices from T are adjacent. The maximum intrinsic stable subset θ is an internally stable subset that is not a proper subset of any other internally stable subset [9], i.e.

$$\forall y_i \in \theta_i (y_i \cap \theta) = \emptyset$$

The concept of a clique is, in fact, the antithesis of an independent subset, since the statement is true:

A click of a graph G - it is a subset of this graph vertices and when it is independent in an additional graph \bar{G} . A subset K' graph G vertices is called a clique in the case when any two vertices in it are adjacent, i.e. if generated sub graph $G(K')$ is complete [10]. In this regard, it is obvious that in order to find the maximum clique in the graph G , it suffices to know the algorithm for finding an internally stable subset in the additional graph. We now turn to the development of a solution algorithm.

Frequently, greedy heuristics are used to form an internally stable subset. They are based on sequential analysis of genes, viewing a series of alternatives, and choosing the best solution. As a rule, they give out local optima, when you first start. However, they have the advantage of high execution speed, of the order of $O(n)$, since, as a rule, the algorithm considers one of the possible options.

In a simplified form, the algorithm will look as follows:

1. Select the top of the graph G with the maximum degree of adjacency and put in θ_1 .
2. The remaining graph G vertices split into two subsets: 1st - vertices adjacent to vertices from θ_1 , 2nd - vertices non-adjacent with vertices from θ_1 .

The choice of the next vertex is made from the 2nd subset (p. 2) with the greatest degree of adjacency.

The general scheme of the modified genetic algorithm is shown in Figure 2.

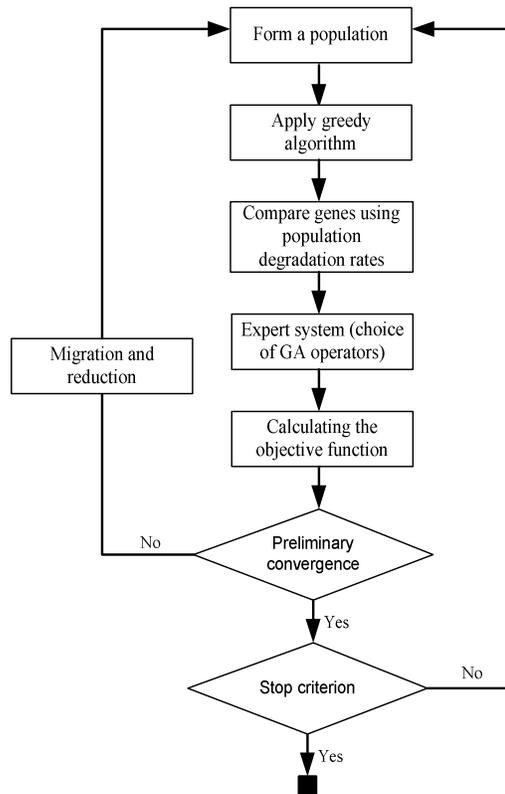


Fig. 2. The general scheme of the modified genetic algorithm

How to compare genes using population degradation indicators is presented in [11]. Based on the decision maker analysis or automatically the genetic algorithm operators are selected: crossover, mutation, transposition, inversion in order to increase the diversity of the population.

In this case, the algorithm used the ordering crossover and greedy crossover (the strategy is described above). The principle of the ordering operation is as follows: a break point is selected by a random method. The left part of each chromosome to the point of break is copied unchanged to the individual of the descendant, starting with the first gene. And the right side is reordered in such a way that the chromosomes of the descendants do not have repeating vertices (Fig. 3).

2	3 (break point)	4	1	5	6	Parent 1
---	--------------------	---	---	---	---	----------

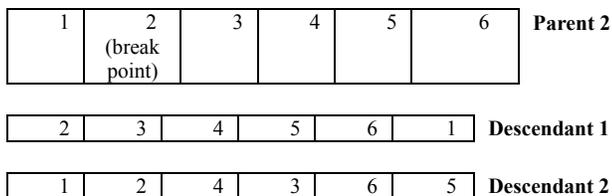


Fig. 3. The principle of the ordering crossover

The initial population is formed using the shotgun strategy, i.e. the population is filled with randomly formed chromosomes.

Selection of parents (the operation of selecting chromosomes for crossover) occurs randomly. Mutation is set when setting up the algorithm.

III. COMPUTER IMPLEMENTATION AND EXPERIMENTAL RESULTS

A series of experiments was carried out to construct an internally stable subset of a graph (ISSG) and a click (CG). In this case, in the first case, we used the ordering crossover operator (OC), and in the second case, the greedy (GC). The calculations were performed on the Amazon Web Services cloud computing infrastructure. For this, a virtual machine (EC2 instance) t3.2xlarge (CPU: 8 and RAM: 32 Gb) was rented. The renting cost was \$ 0.3328 per hour. This significantly reduces the project cost (there is no need to buy a server, hourly payment). The use of cloud services also makes it possible to scale up computing power both vertically (increase the power of one machine) and horizontally (use several machines for parallel computing).

The research results are presented in Table 1. The graph vertices number is $n = 1000$, the number of edges is 700. The number of iterations (NI) is indicated in the 1st column, the population size (PS) - in the 2nd. In the 3rd - the probability of mutation (PM), in the 4th and 6th - the objective function (the ratio of the number of defined and existing internally stable subsets of the graph) for the ISSG task and with 2 different crossovers: OC, GC. In the 5th and 7th columns - the solution time in seconds for the ISSG task and with 2 different crossovers: OC, GC. In the 8th and 10th - the objective function (the ratio of the number of defined and existing of the graph click) for the CG task and with 2 different crossovers: OC, GC. In the 9th and 11th columns - the solution time in seconds for the CG task and with 2 different crossovers: OC, GC respectively.

TABLE I. THE RESEARCH RESULTS

NI	PS	PM	ISSG + OC		ISSG + GC		CG + OC		CG + GC	
1	2	3	4	5	6	7	8	9	10	11
100	20	0,3	0,6	30	0,7	40	0,7	45	0,74	47
150	20	0,12	0,61	40	0,71	55	0,71	46	0,76	49
200	30	0,3	0,62	50	0,72	58	0,73	48	0,8	53
250	30	0,05	0,65	65	0,73	69	0,74	50	0,81	55
400	40	0,2	0,7	88	0,75	95	0,75	55	0,82	59
500	40	0,2	0,78	95	0,76	100	0,8	70	0,85	72
700	50	0,1	0,79	110	0,78	115	0,8	75	0,86	78
800	60	0,3	0,8	130	0,8	136	0,85	80	0,87	82
900	70	0,3	0,9	148	0,86	154	0,86	100	0,9	106
1000	80	0,2	0,9	160	0,87	168	0,87	120	0,91	124
1500	90	0,1	0,9	197	0,88	200	0,9	300	0,92	310
3000	100	0,1	0,9	210	1	216	0,9	380	0,93	385
4000	100	0,3	0,9	270	1	275	0,91	430	0,94	445
5000	100	0,2	0,9	330	1	336	0,91	510	0,94	520
6000	100	0,2	0,9	420	1	427	0,91	570	0,94	590

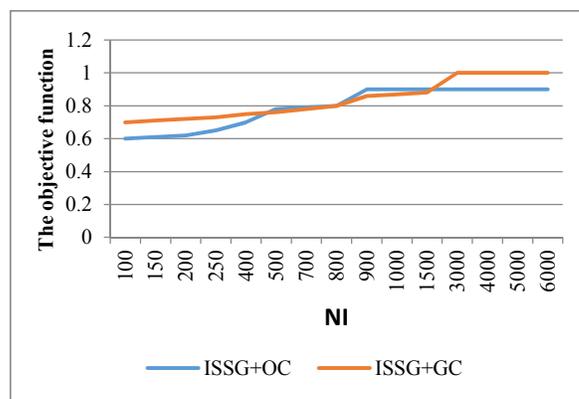


Fig. 4. The graph of the objective function for the ISSG task

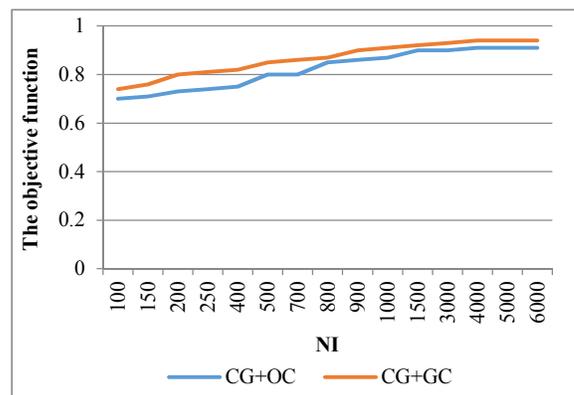


Fig. 5. The graph of the objective function for the CG task

Based on the research results, you can see from the graph on (Fig. 4), the algorithm with the greedy crossover operator shows saturation by 3000 iterations, while the algorithm with the ordering crossing over shows the saturation already at 900

iterations. Practically the same tendency is observed in the clique construction algorithm (Fig. 5). The computational complexity of this algorithm almost linearly depends on the number of iterations and, at best, will be $O(n^2 \log n)$, at worst $O(n^4)$.

IV. CONCLUSION

Currently, the problem of clicks widely has numerous applications. In bioinformatics, it is used in the analysis of genomic databases. In social networks, it is used in data clustering — when various communities are divided into groups (clusters) with common properties. Clustering allocation allows each of them to be processed by a separate auxiliary server. In chemistry, this task underlies the search for the maximum common substructure in the column describing the structure of a chemical compound. At the same time, the amount of input data is huge (input graphs can contain up to a million vertices). The problem of finding extreme subsets in a graph is also reflected in practice when designing various devices and personal computers. Thus, the current area of research is the development of new approaches to finding solutions to these two tasks using advanced IT technologies - cloud computing.

To solve the problem of finding extreme subsets in a graph and building cliques, a modified genetic algorithm was developed. The principal difference of the proposed method is that it considers the degree of population degradation and then increases the diversity of the population through the use of various genetic operators in the evolutionary adaptation block. Several experiments were also conducted, during which a comparison of the developed algorithms with various crossovers was carried out: ordering and greedy. The computational complexity of this algorithm almost linearly depends on the number of iterations and, at best, will be $O(n^2 \log n)$, at worst $O(n^4)$.

The practical research results almost coincided with the theoretical background.

ACKNOWLEDGMENT

The article was supported by The Russian Foundation for Basic Research project № 18-07-00050 and government order 25537.2017/6.7

REFERENCES

[1] Payusova T. I., "ANALYSIS OF THE PROTECTION OF CLOUD SYSTEMS WITH THE HELP OF A GENETIC ALGORITHM", Security of the Information Space: Proceedings of the XII All-Russian Scientific and Practical Conference of Students, Postgraduates and Young

Scientists, Yekaterinburg, 2-4 December 2013 - Yekaterinburg: Ural Publishing House University, 2014, pp. 183-190.

http://elar.urfu.ru/bitstream/10995/46457/1/bip_2014_35.pdf

- [2] Translation Vilchinsky N., "From data storage to information management", Peter, 2016, 544 p.
<https://www.piter.com/collection/informatsionnye-sistemy/product/ot-hraneniya-dannyh-k-upravleniyu-informatsiy-2-e-izdanie>
- [3] Kureychik V.M., "Evolutionary methods for constructing cliques and independent sets of graphs", News of TSURE, 2003, pp. 66-71.
- [4] Reinhold E., Nivergelt U., Deo N., "Combinatorial algorithms, theory and practice", Translated from English EP Lipatova, Mir Publishing House, Moscow 1980, pp. 390, 476 p.
- [5] John H. Holland, "ADAPTATION IN NATURAL AND ARTIFICIAL SYSTEMS", The University of Michigan Press, Ann Arbor, 1992, 232 pp.
- [6] Kureychik V.M., "Genetic algorithms. State, problems, prospects", News of the Academy of Sciences. Theory and control systems, 1999, № 1. pp. 144-160.
- [7] Mukhacheva A.S., Chiglintsev A.V., "Genetic algorithm for finding the minimum in problems of two-dimensional Hamilton cutting", Information technologies, 2001, № 3. pp. 27-31.
- [8] Lipnitsky A.A., "Application of genetic algorithms to the problem of placing rectangles", Cybernetics and Systems Analysis, 2002, № 6, p. 180-184.
- [9] Kofman A., "Introduction to applied combinatorics", Translated from French by V. Myakishev, V. Tarakanova, Mir publishing house, Moscow 1975, p. 180, 479 pp.
- [10] Emelichev V.A., Melnikov O.I., Sarvanov V.I., Tyshkevich R.I., "Lectures on graph theory: Study Guide", Ed. 2nd, rev. Moscow, Book House LIBROCOM, 2009, p. 112, 392 pp.
- [11] Kureichik V.M., Logunova Yu.A., "Analysis of the prospects for the use of the genetic algorithm in solving the traveling salesman problem", INFORMATION TECHNOLOGIES, № 11. Volume 24, 2018, p. 691-697.

A Template Model of Junction Field-Effect Transistors for a Wide Temperature Range

Alexandr M. Pilipenko

Department of Fundamentals of Radio
Engineering
Southern Federal University
Taganrog, Russia
ampilipenko@sfnu.ru

Vadim N. Biryukov

Department of Fundamentals of Radio
Engineering
Southern Federal University
Taganrog, Russia
vnbiryukov@yandex.ru

Nikolay N. Prokopenko

Department «Information Systems and Radio
Engineering»
Don State Technical University
Rostov-on-Don, Russia
prokopenko@sssu.ru

Abstract—The possibility of using a template model for approximation of I-V characteristics of junction field-effect transistors (JFETs) in a wide temperature range (−200 ... 20 °C) is considered. The template model is intended for JFETs which are used in radiation-hardened integrated circuits for processing the signals from sensors of various physical quantities. The template model creation is made by the replacement of one or more parameters of the known physical JFETs model by the relations of power series of control voltages. The template model comprises all physical parameters of the initial JFETs model. The number of additional parameters in the template model is small (no more than four), which makes it possible to use standard minimum search algorithms for parametric identification. The obtained results show that the proposed template model provides the decrease of the JFETs I-V characteristics modeling error at least three times in comparison with the initial physical model independently of the measurements temperature.

Keywords—model, field-effect transistor, template, method of least squares, parametric identification

I. INTRODUCTION

Junction field-effect transistors (JFETs), as well as metal-oxide-semiconductor field-effect transistors (MOSFETs), are used in radiation-hardened integrated circuits (ICs) for processing the signals from sensors of various physical quantities [1]. JFETs have a minimum level of intrinsic noise; MOSFETs have the most advanced technology of mass production [1], [2].

Radiation-hardened ICs are necessary for robotics and space instrument engineering. These ICs can operate in a wide temperature range (−200 ... 20 °C). Field-effect transistors (FETs) which operate at low temperatures have a sufficiently long channel length (1 ... 20 μm) [3], therefore the use of modern models for these transistors does not guarantee a high modeling accuracy [4]. Moreover, a model which accurately approximates FETs I-V characteristics at room temperatures cannot reproduce FETs low temperature characteristics as accurately [5].

Thus, development of FETs models which provide valid results of simulation of radiation-hardened integrated circuits is the actual problem of robotics and space instrument engineering.

The use of template models allows providing an acceptable accuracy of FETs I-V characteristics approximation for helium temperatures [6] and [7]. To create a template model it is necessary to replace one or more parameters of the initial physical model by the function of control voltages or currents [8].

The aim of this paper is to prove that the template models are efficient for JFETs I-V characteristics approximation in a wide temperature range (−200 ... 20 °C).

To achieve the abovementioned aim the following problems are solved in the paper:

- description of known physical models and JFETs template models;
- parametric identification of JFETs models in a wide temperature range;
- determination of dependencies of JFETs models parameters and modeling errors upon temperature.

II. DESCRIPTION OF MODELS

The SPICE-model of FETs, also known as the Shichman-Hodges model, has the following form [9]

$$I(V_{DS}, V_{GS}) = \begin{cases} 0, & \text{for } V_G \leq 0; \\ \beta(2V_G - V_{DS})V_{DS} (1 + \lambda V_{DS}), & \text{for } V_{DS} \leq V_G; \\ \beta V_G^2 (1 + \lambda V_{DS}), & \text{for } V_{DS} > V_G, \end{cases} \quad (1)$$

where I is the drain current; V_{DS} is the drain-to-source voltage; V_{GS} is the gate-to-source voltage; $V_G = V_{GS} - V_{TH}$ is the effective gate voltage; V_{TH} is the threshold voltage; β is the transconductance coefficient; λ is the channel length modulation factor.

The parameters β , V_{TH} , λ of the model (1) are defined using direct measurements, so this model can be used as the initial physical model to create a FETs template model. The technique of template models creation is described in [6] – [8]. In this paper we propose to create a template model by replacing the parameters β and λ of the initial physical model (1) by the following expressions:

$$\beta = \beta_0 \frac{\beta_1 V_G}{\beta_1 V_G + \beta_2 V_G^2}; \quad (2)$$

$$\lambda = \lambda_0 \frac{\lambda_1 V_{DS}}{\lambda_1 V_{DS} + \lambda_2 V_{DS}^2}. \quad (3)$$

where β_0 и λ_0 are the physical parameters of the template model; β_1 , β_2 and λ_1 , λ_2 are the empirical coefficients.

The number of additional empirical coefficients in the obtained template model is four, which makes it possible to use standard minimum search algorithms for parametric identification.

It should be noted that the model (1) is effective as the initial physical model for JFETs. We recommend using the physical model, which comprises the effects of carriers velocity saturation, as the initial model for MOSFETs [7], [10], [11].

III. ALGORITHM OF PARAMETRIC IDENTIFICATION

The parameters of FETs models are determined by the least squares method using the objective function [12]:

$$S = \sum_{k=1}^N \left[\frac{I(V_{DSk}, V_{GSk}) - I_k}{I_k} \right]^2, \quad (4)$$

where N is the number of the experimental points; I_k are the measured values of the drain current; V_{DSk} and V_{GSk} are the measured values of the drain-to-source voltage and the gate-to-source voltage respectively; $I(V_{DSk}, V_{GSk})$ are the drain current values calculated using a FETs model at $V_{DS} = V_{DSk}$ and $V_{GS} = V_{GSk}$.

To solve the problem of the objective function minimum search we used the modified algorithm of Levenberg–Marquardt [13]. To increase the speed of the objective function minimum search the parameters of the initial physical model were used as the initial conditions for the physical parameters of the template model. The initial values of the empirical coefficients β_1 and λ_1 were chosen equal to 1, β_2 and λ_2 – equal to 0.

The accuracy of modeling was estimated using two types of errors described below.

1. The relative root-mean-square (RMS) error:

$$\sigma = \sqrt{\frac{S_{\min}}{N}}, \quad (5)$$

where S_{\min} is the minimum value of the objective function.

2. The maximum relative error of the model:

$$\delta_{\max} = \max |\delta_k|, \quad (6)$$

where $\delta_k = [I(V_{DSk}, V_{GSk}) - I_k] / I_k$ is the relative error of the model at each point of the I-V characteristic.

IV. RESULTS OF MODELING

Table 1 shows the results of the JFET modeling for several temperatures in the range $-200 \dots 20$ °C. In this work the p-channel JFET made on the radiation-hardened analog gate array ABMK 1-3 was considered as the test sample for research. The gate array ABMK 1-3 is produced by JSC "INTEGRAL" (Minsk, Belarus) with 1.5 microns design rule and is intended for the fabrication of low-noise and broadband analog ICs [1].

TABLE I. RESULTS OF PARAMETRIC IDENTIFICATION

Model	Model parameters							Model errors	
	β_0 [mA/V ²]	V_{TH} [V]	λ_0 [V ⁻¹]	β_1 [V ⁻¹]	β_2 [V ⁻²]	λ_1 [V ⁻¹]	λ_2 [V ⁻²]	σ [%]	δ_{\max} [%]
<i>T = 20 °C</i>									
(1)	2.273	2.020	0.0338	-	0	-	0	4.5	15.3
(1)&(2)&(3)	2.663	1.946	1.1594	-0.436	-0.0976	0.0150	0.0107	1.2	3.4
<i>T = -60 °C</i>									
(1)	3.652	1.874	0.0283	-	0	-	0	6.2	16.4
(1)&(2)&(3)	5.044	1.802	0.9846	0.157	0.0579	0.0147	0.00957	1.5	5.0
<i>T = -120 °C</i>									
(1)	4.660	1.752	0.0295	-	0	-	0	8.2	19.9
(1)&(2)&(3)	6.675	1.698	1.0383	0.399	0.190	0.0233	0.0158	2.4	6.2
<i>T = -200 °C</i>									
(1)	1.477	1.580	0.0732	-	0	-	0	9.3	27.4
(1)&(2)&(3)	1.666	1.439	1.6092	-2.297	-0.494	0.124	0.0835	2.7	6.9

The model (1) was chosen as the initial physical JFET model for parametric identification. The template model (1)&(2)&(3) was obtained on the basis of the model (1) by replacing the parameters β and λ by the expressions (2) and (3) respectively. It should be noted that the initial physical model (1) is the special case of the template model (1)&(2)&(3) with $\beta_2 = \lambda_2 = 0$ or $\beta = \beta_0, \lambda = \lambda_0$.

As we can see from Table 1 the use of the template model provides the decrease of the RMS and maximum errors of JFET modeling approximately in 3 – 4 times in comparison with the initial physical model independently of the operating temperature.

Fig. 1 illustrates the measured I-V characteristics of JFET, as well as the I-V characteristics calculated using the initial and template models. The results shown in Fig. 1 confirm the effectiveness of the template model use in the whole operating area of the I-V characteristics.

Fig. 2 shows the dependencies of the JFET transconductance coefficient and threshold voltage upon temperature. As we can see from Fig. 2 the type of the temperature dependencies of the corresponding parameters for the initial and the template models differ slightly.

The dependencies of the JFET models errors upon temperature are shown in Fig. 3. As we can see from Fig. 3 the JFET modeling errors increase with the temperature decreasing. The RMS and the maximum errors of the initial model converge to 10 % and 30 % respectively, at the same time the similar errors of the template model do not exceed 3 % and 7 % respectively.

V. CONCLUSIONS

The efficiency of the template model use for JFETs I-V characteristics approximation in a wide temperature range ($-200 \dots 20 \text{ }^\circ\text{C}$) is proved in this paper. The template model provides the increase of JFETs I-V characteristics modeling accuracy in 3 – 4 times in comparison with the initial physical model. The proposed technique of template models creation is also applicable for MOSFETs [6], [7]. It should be noted that it is required to use different initial physical models for different types of FETs, because for the more accurate initial model is the more accurate the template model.

The proposed algorithm of parametric identification is realized on the basis of standard methods for the objective function optimization. This algorithm allows obtaining the error of determining the parameters of JFETs models comparable to the error of I-V characteristics measurement.

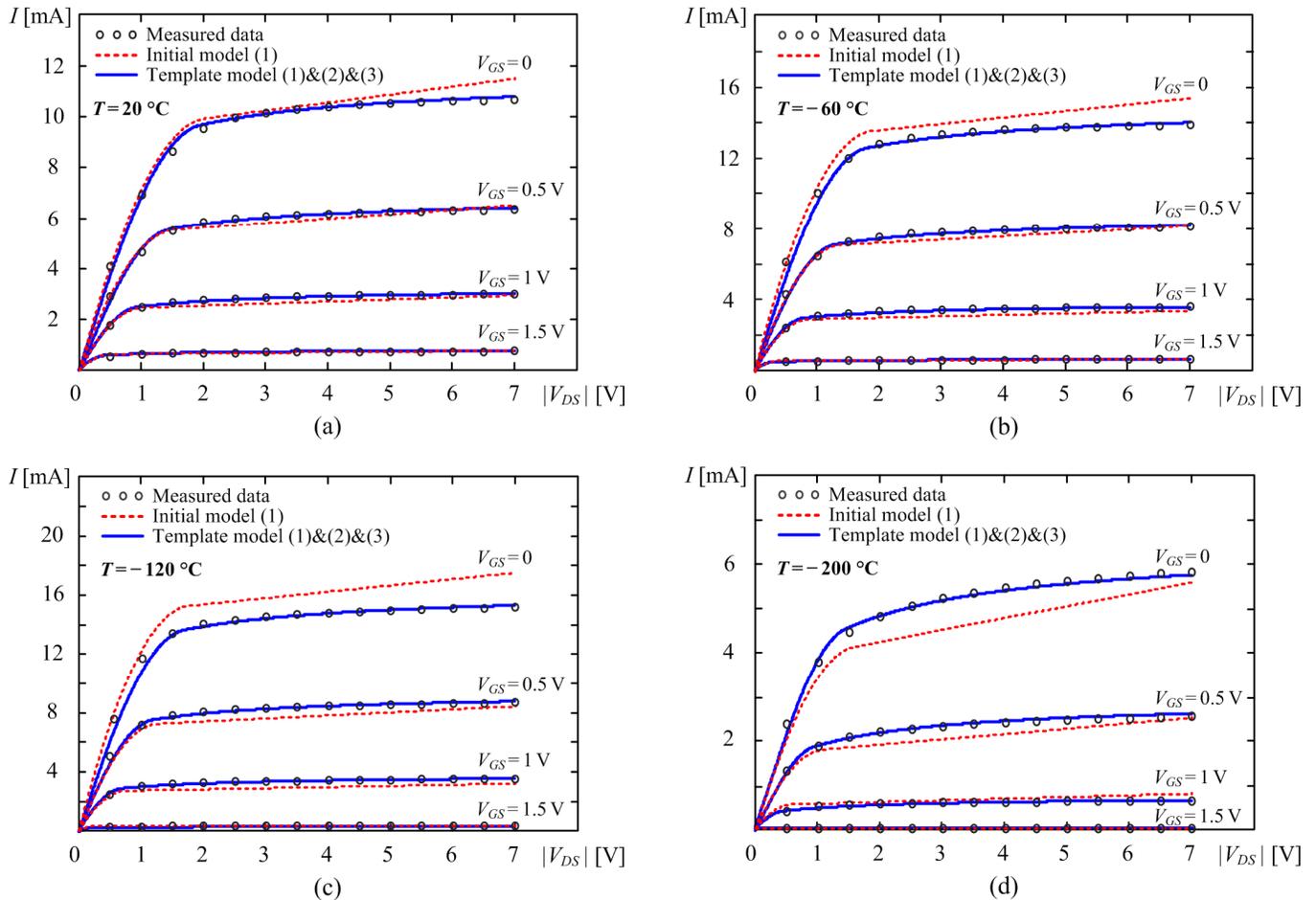


Fig. 1. Measured and calculated I-V characteristics of JFET under $T = 20 \text{ }^\circ\text{C}$ (a), $T = -60 \text{ }^\circ\text{C}$ (b), $T = -120 \text{ }^\circ\text{C}$ (c), $T = -200 \text{ }^\circ\text{C}$ (d).

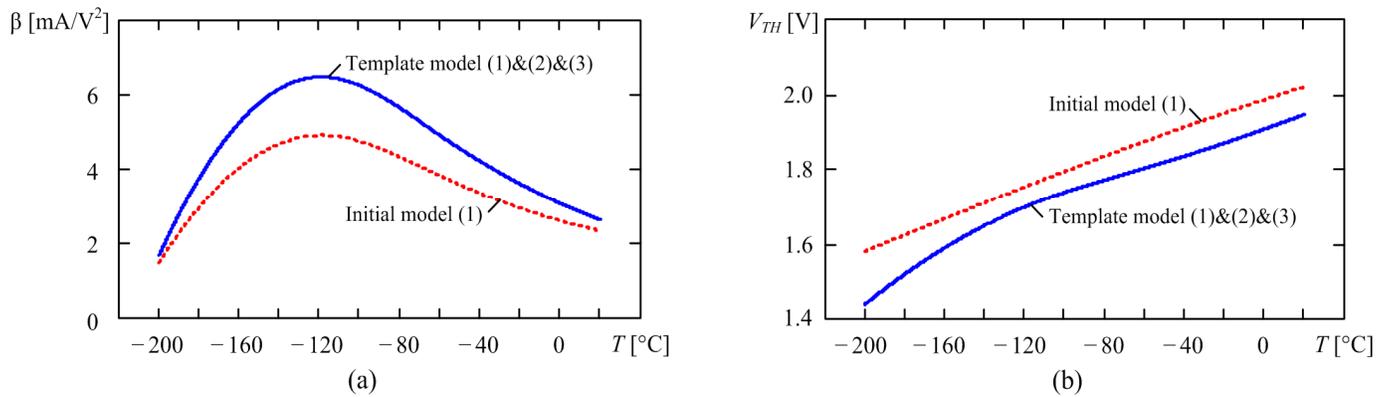


Fig. 2. Dependencies of the JFET transconductance coefficient (a) and threshold voltage (b) upon the temperature.

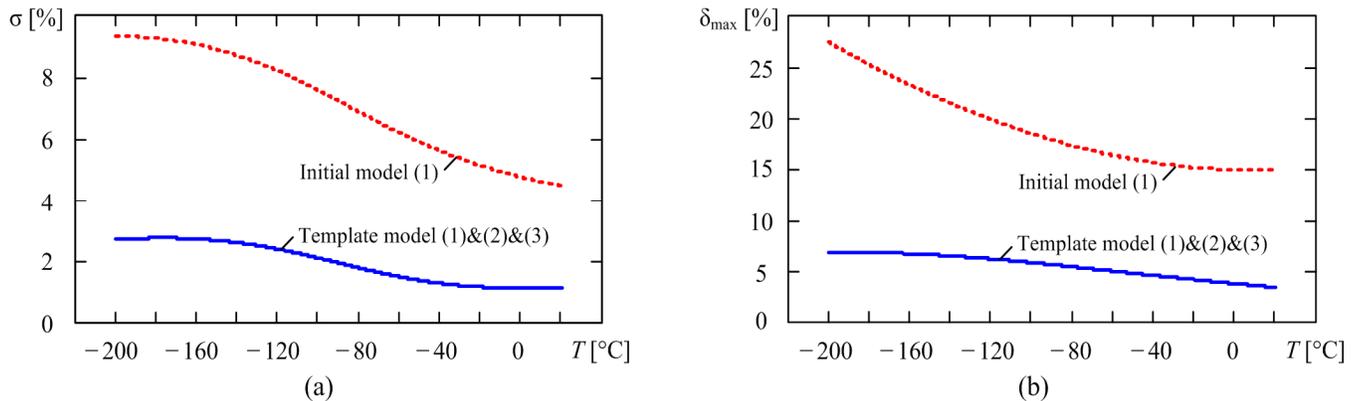


Fig. 3. Dependencies of the JFET models RMS (a) and maximum (b) errors upon the temperature.

The template model comprises all physical parameters of the initial model. Temperature dependencies of the JFET transconductance coefficient and threshold voltage for the initial and the template models differ slightly, so the template model can be used to estimate JFETs base parameters in a wide temperature range.

REFERENCES

- [1] O. V. Dvornikov, V. L. Dzatlau, N. N. Prokopenko, K. O. Petrosiants, N. V. Kozhukhov, and V. A. Tchekhovski, "The accounting of the simultaneous exposure of the low temperatures and the penetrating radiation at the circuit simulation of the BiJFET analog interfaces of the sensors," 2017 International Siberian Conference on Control and Communications (SIBCON). Proceedings, 2017, doi:10.1109/SIBCON.2017.7998507.
- [2] S. S. Li, *Semiconductor Physical Electronics*, 2nd ed. Springer, 2006.
- [3] H. Nagata, H. Shibai, T. Hirao, T. Watabe, M. Noda, Y. Hibi, M. Kawada, and T. Nakagawa, "Cryogenic Capacitive Transimpedance Amplifier for Astronomical Infrared Detectors," *IEEE Transactions on Electron Devices*, vol. 51, no. 2, pp. 270–278, February 2004, doi: 10.1109/TED.2003.821764.
- [4] Y. P. Tsvividis and K. Suyama, "MOSFET modeling for analog circuit CAD: Problems and prospects," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 3, pp. 210–216, March 1994, doi: 10.1109/4.278342.
- [5] J. Kan, S. Weifeng, and S. Longxing, "A sub-circuit MOSFET model with a wide temperature range including cryogenic temperature," *Journal of Semiconductors*, vol. 32, no. 6, June 2011, doi: 10.1088/1674-4926/32/6/064002.
- [6] V. N. Biryukov and A. M. Pilipenko, "Measurement-Based MOSFET Model for Helium Temperatures," *Proceedings of 2015 IEEE East-West Design and Test Symposium (EWDTS)*, 2015, pp. 241-244, doi: 10.1109/EWDTS.2015.7493109.
- [7] V. N. Biryukov, "Template modeling of a p-channel MOSFET," *Zhurnal Radioelektroniki – Journal of Radio Electronics*, no. 2, February 2019. doi: 10.30898/1684-1719.2019.2.11.
- [8] S. Van den Bosch and L. Martens "Approximation of State Functions in Measurement-Based Transistor Model," *IEEE Transactions on Microwave Theory and Techniques*, vol. 47, no. 1, pp. 14-17, January 1999, doi: 10.1109/22.740069.
- [9] G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE*, 2nd ed. McGraw-Hill, 1993.
- [10] L. Bisdounis, S. Nikolaidis, O. Koufopavlou, and C. E. Goutis. "Switching response modeling of the CMOS inverter for sub-micron devices," *Proceedings of the conference on Design, automation and test in Europe*, 1998, pp. 729-737.
- [11] M. Shoji, *Theory of CMOS Digital Circuits and Circuit Failures*. Princeton, New Jersey: Princeton University Press, 1992.
- [12] A. M. Pilipenko and V. N. Biryukov, "Modeling of MOSFETs Parameters and Volt-Ampere Characteristics in a Wide Temperature Range for Low Noise Amplifiers Design," *Proceedings of IEEE East-West Design & Test Symposium (EWDTS)*, 2014, pp. 156-159, doi: 10.1109/EWDTS.2014.7027065.
- [13] A. M. Pilipenko and V. N. Biryukov, "Efficiency improvement of the random search algorithm for parametric identification of electronic components models," 2016 International Siberian Conference on Control and Communications (SIBCON). Proceedings, 2016, doi: 10.1109/SIBCON.2016.7491703.

Synthesis of Signal Quadrature Processing Algorithms under the Influence of Band-limited non-Gaussian Noise

Vladimir Mikhaylovich Artyushenko
Information technology and management systems department
Technological university
Korolev city, Russian Federation
artuschenko@mail.ru

Vladimir Ivanovich Volovach
Informational and electronic service department
Volga region state university of service
Togliatty city, Russian Federation
volovach.vi@mail.ru

Abstract—The synthesis of quasi-optimal algorithms for processing useful signals under the influence of additive non-Gaussian noise with a band-limited spectrum using quadrature processing is carried out. These algorithms use methods of nonlinear Markov filtering. Two practically important cases when the correlated and independent samples of influencing noise quadrature components are accounted. For these cases, the block diagrams processing signals under the influence of band-limited noise, including those arising from the reflection from the underlying surface and concentrated reflectors, are obtained. It is shown that the synthesized algorithms implement nonlinear processing of quadratures of the acting non-Gaussian noise. It is noted that the characteristic of the nonlinear conversion unit of noise quadratures is described by two-dimensional transition of probability density function of additive noise quadratures. It is shown that the extrapolated estimate leads to a further simplification of the recurrence equations. The nonlinearity of the channels is due to the nonlinear dependence of the information sequence of the signal and the non-Gaussian noise. The case in which a stationary mode is set in the synthesized demodulator is described.

Keywords—probability density function, band-limited non-Gaussian noise, quadrature processing, nonlinear Markov filtering, quasi-optimal demodulation algorithms.

I. INTRODUCTION

Many scientific articles [1-4, etc.] consider the processing of useful signals under the influence of noise with a limited range of spectrum. The purpose of these studies is to detect a useful signal or measure its parameters. In these publications it is assumed that this noise is a Gaussian process, the envelope of which is generally described by the Rice probability density function (PDF).

At the same time, in practice, it is often necessary to investigate the effect of non-Gaussian band-limited noise with a PDF envelope (amplitude) not subject to Rice distribution on demodulation (filtering) of useful signals, often having small amplitude. An example of the above noise is the noise in radio communication, radio detection and sonar detection which is due to both the work of interfering neighboring radio stations and the multipath nature of the received signal at the input of the demodulator.

Earlier in a number of works [5-8, etc.] were considered issues of processing (detection) of useful signals, which are affected by non-Gaussian noise. However, only broadband noise which can be represented by independent samples of quadrature components is studied there.

In the course of this work, it is necessary to develop quadrature algorithms for processing useful signals, which are affected by non-Gaussian noise with band-limited

spectrum. In order to do that quadrature processing is used. The width of the spectrum of the considered noise can be almost the same as a band of a useful signal or it can be wider or narrower. Note that such noise occurs in radio detecting and sonar detecting due to reflection from the underlying surface and concentrated reflectors.

Using the theory of signal detection under the influence of noise, including non-Gaussian distribution [9-11, etc.], as well as methods of nonlinear Markov filtering, allows to obtain quasi-optimal algorithms for demodulation (filtering) of signals under the influence of band-limited noise, PDF of which differs from Gauss function. In order to do that we use quadrature processing, especially when the adjacent samples of the noise are correlated with each other, i.e. when it is a general case.

II. GENERAL CASE OF QUADRATURE SIGNAL PROCESSING UNDER THE INFLUENCE OF ADDITIVE NON-GAUSSIAN NOISE

Let the input of the demodulating (receiving) block be affected by the signal and non-Gaussian noise, and their mixture is a narrow-band oscillation

$$y(t) = U_e(t) \cos[\omega_c t + \Theta_e(\lambda, t)], \quad (1)$$

where $U_e(t)$ is the envelope of the input mixture, λ is the parameter of information process, $\Theta_e(\lambda, t) = \varphi_e(\lambda, t) + \varphi_{e.o.}$, ω_c represents the carrier frequency, $\varphi_e(\lambda, t)$ this is the phase of the input mixture, carrying the information about the demodulated (filtered) sequence, $\varphi_{e.o.}$ is a random phase of the input mixture.

We get such a model, for example, when the band-limited signal $s(\lambda, t)$ is exposed to band-limited noise $n(t)$.

Let's assume that the narrow-band oscillation $y(t)$ (1) is the additive mixture of the signal $s(\lambda, t)$ containing the information process $\lambda(t)$ and additive noise $n(t)$

$$y_h = s(\lambda, t_h) + n_h,$$

moreover, range of useful signal is more narrow-band than spectrum noise.

We assume that the signal $s(\lambda, t)$ is narrow-band with a random phase $\varphi_{s.o.}$. Generally, it is modulated both in amplitude and phase, so that

$$s(\lambda, t) = U_s(t) \cos[\omega_c t + \Theta_s(\lambda, t)],$$

where $U_s(t)$ is the envelope of the narrow-band signal; $\Theta_s(\lambda, t) = \Theta_c(\lambda, t) + \varphi_{s.o.}$, $\Theta_c(\lambda, t)$ is the phase of the signal which carries information about the demodulated information process $\lambda(t)$. It is usually the central frequency of the spectrum.

This additive noise is described by equality

$$n(t) = U_{an}(t) \cos[\omega_c t + \Theta_{an}(t)],$$

where $U_{an}(t)$ is the envelope of noise; $\Theta_{an}(t) = \varphi_{an}(t) + \varphi_{an.o}$; $\varphi_{an}(t)$ and $\varphi_{an.o}$ are regular and random phases of additive noise respectively.

Taking into consideration that the input mixture $y(t)$ is a narrow-band process, we present it as a quadrature transformation:

$$y(t) = U_{1.e}(t) \sin \omega_c t + U_{2.e}(t) \cos \omega_c t,$$

where $U_{1.e}(t) = U_e(t) \sin(\Theta_e(\lambda, t))$; $U_{2.e}(t) = U_e(t) \cos(\Theta_e(\lambda, t))$ are quadrature components of a complex envelope of the input mixture.

The quadrature components $U_{1.e}(t)$ and $U_{2.e}(t)$ can be presented as a sum:

$$U_{1.e}(t) = U_{1.si}(\lambda, t) + U_{1.an}(t);$$

$$U_{2.e}(t) = U_{2.si}(\lambda, t) + U_{2.an}(t),$$

where $U_{1.si}(\lambda, t)$, $U_{2.si}(\lambda, t)$; $U_{1.an}(t)$, $U_{2.an}(t)$ are quadrature components, respectively, of the signal that contains the information about the demodulated process and additive noise gradually changing as compared with $\cos \omega_c t$.

Note, that when the frequency ω_c is known, the gradually changing processes $U_{1.e}(t)$ and $U_{2.e}(t)$ are sufficient to present the statistics of observation.

Within the observation interval T sample values $y(t_h) = y_h$ are taken from the input mixture at the interval of T_0 .

Then the observation equation (1) will be:

$$\begin{aligned} y(t_h) &= s(\lambda_h, t_h) + n_h = \\ &= U_{1.e}(t_h) \sin \omega_c t_h + U_{2.e}(t_h) \cos \omega_c t_h = \\ &= [U_{1.si}(\lambda_h, t_h) + U_{1.an}(t_h)] \sin \omega_c t_h + \\ &+ [U_{2.si}(\lambda_h, t_h) + U_{2.an}(t_h)] \cos \omega_c t_h, \end{aligned}$$

where $\{\lambda_h\}$, $\{n_h\}$ describe, respectively, demodulated (filtered) information sequence and additive noise; $\{U_{1.e}(t_h)\} = \{U_{1.e.h}\}$ and $\{U_{2.e}(t_h)\} = \{U_{2.e.h}\}$ describe the sequence of quadrature components of the complex envelope of the input mixture; $\{U_{1.si}(\lambda_h, t_h)\} = \{U_{1.si}(\lambda_h)\}$, $\{U_{2.si}(\lambda_h, t_h)\} = \{U_{2.si}(\lambda_h)\}$, $\{U_{1.an}(t_h)\} = \{U_{1.an.h}\}$, $\{U_{2.an}(t_h)\} = \{U_{2.an.h}\}$ describe sequences respectively of quadrature components of envelopes of the useful signal and additive noise.

Given that $\{k_h\}$, where $k = \lambda, n$ form a Markov process with known statistical characteristics which satisfies the expression

$$\begin{aligned} k_h \in W_k(k^h) &= W_k(k_0, \dots, k_h, t_0, \dots, t_h) = \\ &= W_k(k_h, t_h | k_{h-1}, t_{h-1}) W_k(k_{h-1}, t_{h-1} | k_{h-2}, t_{h-2}) \dots \\ &\dots W_k(k_1, t_1 | k_0, t_0) W_k(k_0, t_0), \end{aligned}$$

where $W_k(k_0, t_0)$ is the initial one-dimensional PDF of the considered sequence $\{k_h\}$; k_0 is the value of the considered sequence $\{k_h\}$ at the initial moment t_0 .

Two-dimensional sequence $\{U_{1.d,h}, U_{2.d,h}\}$, where $d = y, s, n$, also forms a Markov process with known statistical characteristics

$$\begin{aligned} \{U_{1.d,h}, U_{2.d,h}\} \in W_{U_1 U_2} \left\{ (U_{1,d}, U_{2,d})^h \right\} = \\ = W_{U_1 U_2} \left\{ (U_{1,d,h}, U_{2,d,h}), t_h | (U_{1,d,h-1}, U_{2,d,h-1}), t_{h-1} \right\} \times \\ \times W_{U_1 U_2} \left\{ (U_{1,d,h-1}, U_{2,d,h-1}), t_{h-1} | (U_{1,d,h-2}, U_{2,d,h-2}), t_{h-2} \right\} \dots \\ \dots W_{U_1 U_2} \left\{ (U_{1,d,1}, U_{2,d,1}), t_1 | (U_{1,d,0}, U_{2,d,0}), t_0 \right\} \times \\ \times W_{U_1 U_2} \left\{ (U_{1,d,0}, U_{2,d,0}), t_0 \right\}, \end{aligned}$$

where $W_{U_1 U_2} \left\{ (U_{1,d,0}, U_{2,d,0}), t_0 \right\}$ is a two-dimensional initial PDF of the considered sequence of quadratures $\{U_{1,d,h}, U_{2,d,h}\}$; $(U_{1,d,0}, U_{2,d,0})$ are the values of the considered sequence $\{U_{1,d,h}, U_{2,d,h}\}$ at the moment t_0 .

A priori, we assume that the transition PDF $W_k(k_h | k_{h-1})$, $W_{U_1 U_2} \left\{ (U_{1,d,h}, U_{2,d,h}), t_h | (U_{1,d,h-1}, U_{2,d,h-1}), t_{h-1} \right\}$ and the initial PDF $W_{U_1 U_2} \left\{ (U_{1,d,0}, U_{2,d,0}), t_0 \right\}$ are known.

Using the results [9, 11], we analyze the structure of the "final" a posteriori PDF:

$$W_y(\lambda_h) = C_h \int \exp\{L_{n,k}(\lambda_h)\} W_\lambda(\lambda_h | \lambda_{h-1}) W_y(\lambda_{h-1}) d\lambda_{h-1};$$

$$W_y(\lambda_1) = C_1 \exp\{L_{n,k}(\lambda_1)\} W_\lambda(\lambda_1),$$

where C_1 and C_h are constant normings for signal processing in quadratures.

Let the logarithm of the likelihood function exist and can be written as follows

$$L_{n,k} = \ln W_l \left(y_h - s(\hat{\lambda}_h, t_h) | y_{h-1} - s(\hat{\lambda}_{h-1}, t_{h-1}) \right). \quad (2)$$

We assume that the signal/noise ratio in the demodulating device (at its output) is small, then, we write a recurrent ratio for the posterior PDF for the problem under consideration. We obtain recurrent equations by taking a Taylor series expansion around the preliminary estimate (chosen by one of the known methods) and limiting it to linear and quadratic terms in the Gaussian approximation. These equations describe the quadrature algorithm of the demodulated (filtered) information sequence by quadrature components under the action of additive narrow-band noise (with a band-limited spectrum) with generally a non-Gaussian distribution:

$$\hat{\lambda}_h = \hat{\lambda}_h^o + \hat{\sigma}_{\lambda,h}^2 \left[L_{\lambda,h}^{\lambda'} + L_{\lambda,h}^{\lambda''} - N_h \left(L_{\lambda,h-1}^{\lambda'} + L_{\lambda,h-1}^{\lambda''} \right) \right]; \quad (3)$$

$$\hat{\sigma}_{\lambda,h}^2 = \left[L_{\lambda,h,h}^{\lambda''} + L_{\lambda,h,h}^{\lambda''} - N_h \left(L_{\lambda,h,h-1}^{\lambda''} + L_{\lambda,h,h-1}^{\lambda''} \right) \right]^{-1}; \quad (4)$$

$$\begin{aligned} \hat{\lambda}_1 &= \hat{\lambda}_1^o + \hat{\sigma}_{\lambda,1}^2 \left[L_{\lambda,1}^{\lambda'} + L_{\lambda,1}^{\lambda''} \right]; \hat{\sigma}_{\lambda,1}^2 = \left[L_{\lambda,1,1}^{\lambda''} + L_{\lambda,1,1}^{\lambda''} \right]^{-1}; \\ F_h &= \left[L_{\lambda,h,h}^{\lambda''} + L_{\lambda,h,h}^{\lambda''} \right] \left[\hat{\sigma}_{\lambda,h-1}^2 + L_{\lambda,h-1,h-1}^{\lambda''} + L_{\lambda,h-1,h-1}^{\lambda''} \right]^{-1}; \quad (5) \end{aligned}$$

$$L_{\lambda,h}^{\lambda''} = Z_{1.an,h}(\cdot) U'_{1.si}(\hat{\lambda}_h) + Z_{2.an,h}(\cdot) U'_{2.si}(\hat{\lambda}_h);$$

$$L_{\lambda,h-1}^{\lambda''} = Z_{1.an,h-1}(\cdot) U'_{1.si}(\hat{\lambda}_{h-1}) + Z_{2.an,h-1}(\cdot) U'_{2.si}(\hat{\lambda}_{h-1});$$

$$\begin{aligned} L_{\lambda,h,h}^{\lambda''} &= Z'_{1.an,h}(\cdot) \left[U'_{1.si}(\hat{\lambda}_h) \right]^2 + Z'_{2.an,h}(\cdot) \left[U'_{2.si}(\hat{\lambda}_h) \right]^2 - \\ &- Z_{1.an,h}(\cdot) U''_{1.si}(\hat{\lambda}_h) - Z_{2.an,h}(\cdot) U''_{2.si}(\hat{\lambda}_h); \end{aligned}$$

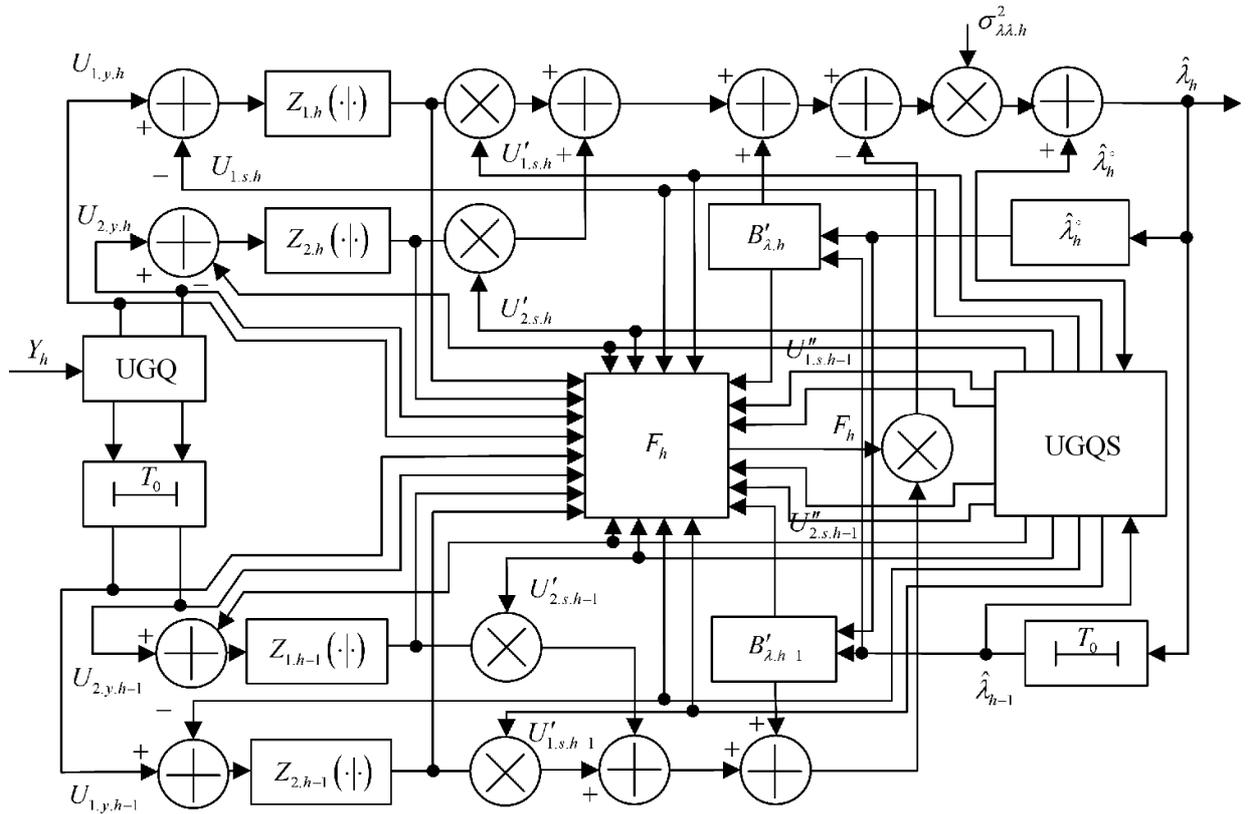


Fig. 1. Block diagram of the information process demodulation channel

$$L_{\lambda,h-1,h-1}^{n*} = Z'_{1,an,h-1}(\cdot) [U'_{1,si}(\hat{\lambda}_{h-1})]^2 + Z'_{2,an,h-1}(\cdot) [U'_{2,si}(\hat{\lambda}_{h-1})]^2 - Z_{1,an,h-1}(\cdot) U'_{1,si}(\hat{\lambda}_{h-1}) - Z_{2,an,h-1}(\cdot) U'_{2,si}(\hat{\lambda}_{h-1});$$

$$L_{\lambda,h-1,h-1}^{n*} = \frac{d}{dU_{1,an,h-1}} Z_{1,an,h}(\cdot) U'_{1,si}(\hat{\lambda}_h) U'_{1,si}(\hat{\lambda}_{h-1}) + \frac{d}{dU_{2,an,h-1}} Z_{2,an,h}(\cdot) U'_{2,si}(\hat{\lambda}_h) U'_{2,si}(\hat{\lambda}_{h-1}),$$

at that

$$Z_{j,an,h-i}(\cdot) = Z_{j,an,h-i}(U_{1,an,h}, U_{2,an,h} | U_{1,an,h-1}, U_{2,an,h-1}) = -\partial W_{an} \ln(U_{1,an,h}, U_{2,an,h} | U_{1,an,h-1}, U_{2,an,h-1}) / \partial U_{j,an,h-i}, \quad (6)$$

$U'_{j,si}(\hat{\lambda}_{h-i}) = dU_{j,si}(\hat{\lambda}_{h-i}) d\hat{\lambda}_{h-i}$, $j = 1, 2$; $i = 0, 1$; $\hat{\lambda}_h^0$ is a preliminary estimate of the information parameter at the step h .

The block diagram presents only the implementation of the algorithm for estimating the demodulated (filtered) sequence for the sake of simplicity. It is shown in Fig. 1, where UGQ is a unit generating quadratures of the input mixture; UGQS is a unit generating quadratures of the useful signal.

The diagram presents a non-linear non-stationary four-channel device where processing in the channels at steps h and $h-1$ is carried out according to the quadratures of the input mixture and the useful signal, formed, respectively, in the blocks UGQ and UGQS.

The structure of each of the channels includes a nonlinear block of noise quadratures (NLB NQ) which provides

suppression of additive noise quadratures influencing the corresponding quadrature of the useful signal. The characteristic of NLB NQ is determined by a two-dimensional transfer PDF of the non-Gaussian the described additive noise quadratures according to the expression (6).

Note that the nonlinearity of the channels is explained by two reasons: nonlinear dependence of the information sequence of the signal $s(\lambda, t)$ and the non-Gaussian nature of the PDF $W(n)$ of the acting additive noise.

III. INFLUENCE OF NON-GAUSSIAN BAND-LIMITED ADDITIVE NOISE WITH INDEPENDENT SAMPLING OF QUADRATURE COMPONENTS ON QUADRATURE SIGNAL PROCESSING

If sampling of quadrature components of additive noise that affects the analyzed signal is independent at h and $h-1$ steps, the expressions (3) and (4) are simplified and take the form:

$$\hat{\lambda}_h = \hat{\lambda}_h^0 + \hat{\sigma}_{\lambda\lambda,h}^2 [L'_{\lambda,h} + L_{\lambda,h}^{n*} - N_h L'_{\lambda,h-1}];$$

$$\hat{\sigma}_{\lambda\lambda,h}^2 = [L'_{\lambda,h,h} + L_{\lambda,h,h}^{n*} - N_h L'_{\lambda,h,h-1}]^{-1}. \quad (7)$$

Thus, the block diagram of the demodulator is greatly simplified (Fig. 2).

The characteristic of the BNT NQ in this case is described by the expression:

$$Z_{j,an,h-i}(U_{j,an,h} | U_{j,an,h-1}) = -\partial W_{an} \ln(U_{j,an,h} | U_{j,an,h-1}) / \partial U_{j,an,h-i};$$

$\varphi = 1, 2$; $i = 0, 1$.

If you choose an extrapolated $\hat{\lambda}_h^0 = \hat{\lambda}_{e,h}$ as a preliminary estimate, the expressions (7) will become even more simplified:

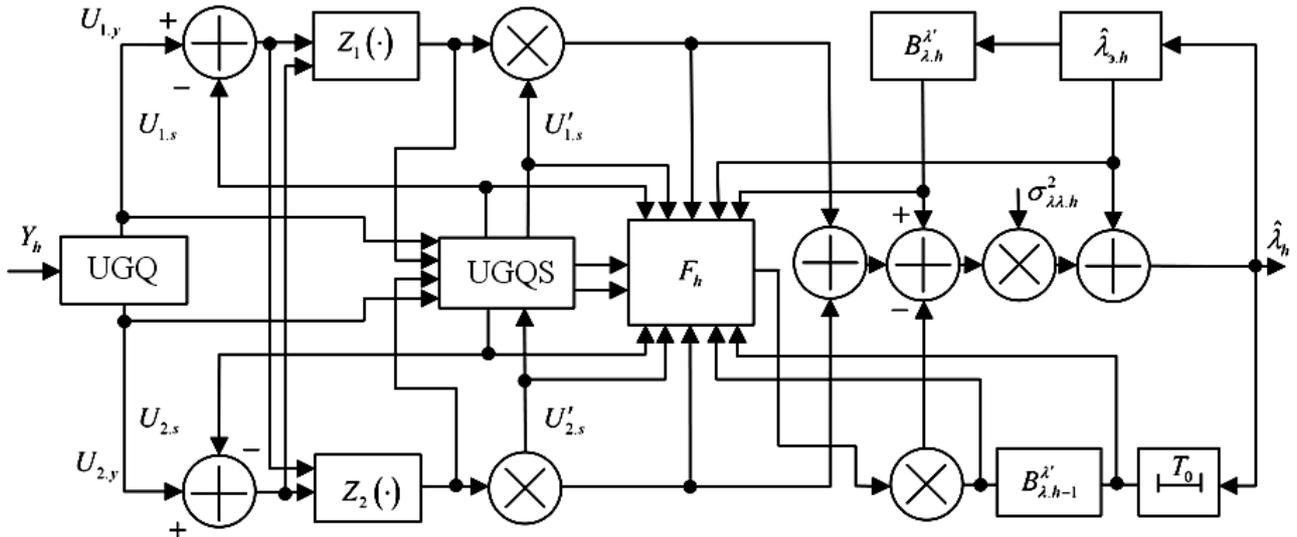


Fig. 2. A simplified block diagram of the information process demodulation channel

$$\hat{\lambda}_h = \hat{\lambda}_{e,h} + \hat{\sigma}_{\lambda,h}^2 L_{\lambda,h}^{n'};$$

$$\hat{\sigma}_{\lambda,h}^2 = [L_{\lambda,h,h}^{\lambda'} + L_{\lambda,h,h}^{n'} - N_h L_{\lambda,h,h-1}^{\lambda'}]^{-1},$$

here N_h is determined in (5).

If the processes $\lambda(t)$ and $n(t)$, and, consequently, the sequences $\{\lambda_h\}$, $\{n_h\}$ are stationary, then a stationary mode is established over time in the synthesized demodulator under certain conditions. For this mode the variance of the estimator of the demodulated (filtered) sequence can be considered more or less time independent. In this case, the quadratures $U_{1,d,h}$ and $U_{2,d,h}$ where $d = y, s, n$ are also stationary, that is $U_{i,d,h} \approx U_{i,d,h}$, where $i = 1, 2$.

The block diagram of the demodulator is simplified by replacing the channel of posterior error estimation with a gain factor. This gain factor characterizes the value of the posterior error variance of the demodulation (filtration) of the information sequence at step h , $\hat{\sigma}_{\lambda,h}^2$ by its stationary value $\sigma_{e,\lambda}^2$.

IV. CONCLUSIONS

Synthesis of the quadrature algorithms for processing narrow-band signals under the influence of band-limited non-Gaussian noise in the dependent and independent adjacent samples can be carried out by nonlinear Markov filtering. Sequences of quadrature components of the envelope of the processed signal and additive non-Gaussian noise together with a two-dimensional sequence form a Markov process with known statistical characteristics. Block diagrams presenting synthesized algorithms are obtained. It is shown that a distinctive feature of the latter is the nonlinear processing of quadratures of generally non-Gaussian noise influencing the signal. The nonlinearity of demodulator channels is due to the nonlinear dependence of the information sequence of the useful signal and the non-Gaussian view the PDF of noise. Characteristic of the block of nonlinear transformations of the quadrature demodulator is defined by two-dimensional transition PDF of the additive noise quadratures. The extrapolated estimate significantly

simplifies the recurrent equations and, consequently, the demodulator diagrams are simplified as well. The conditions for obtaining the stationary mode of the synthesized demodulator are described.

REFERENCES

- [1] Van Trees, K. Bell, and Z. Tiany, Detection Estimation and Modulation Theory, 2nd Edition, Part I, Detection, Estimation, and Filtering Theory. London: Wiley & Sons, Inc., 2013.
- [2] V. P. Tuzlukov, Signal Processing Noise, Boca Raton, London, New York, Washington D.C.: CRC Press, Taylor & Francis Group, 2002.
- [3] Ellingson, Radio System Engineering. Cambridge University Press, 2016.
- [4] M. Barkat, Signal Detection and Estimation. Norwood: Artech House, 2005.
- [5] H S. Kassam, Signal Detection in Non-Gaussian Noise. Berlin: Springer, 1988.
- [6] G. L. Charvat Small and Short-Range Radar Systems. CRC Press, 2014.
- [7] J. Yang, Y. Cheng, H. Wang, Y. Li, and X. Hua, "Unknown stochastic signal detection via non-Gaussian noise modeling," Proceedings 2015 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Ningbo, China, 2015, pp. 1–4. DOI: [10.1109/ICSPCC.2015.7338861](https://doi.org/10.1109/ICSPCC.2015.7338861)
- [8] E. Palahina, and V. Palahin, "Signal detection in additive-multiplicative non-Gaussian noise using higher order statistics", 2016 26th International Conference Radioelektronika (RADIOELEKTRONIKA), IEEE, 2016, pp. 262-267. DOI: [10.1109/RADIOELEK.2016.7477367](https://doi.org/10.1109/RADIOELEK.2016.7477367)
- [9] V. M. Artyushenko, V. I. Volovach, and V. N. Budilov, "Synthesis and analysis of discriminators meter information parameters signal under non-Gaussian noise with band pass spectrum", Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2017). Novi Sad, Serbia, Sept 29-Oct 2, 2017. – Kharkov: KNURE, 2017. P. 355-358. DOI: [10.1109/EWDTS.2017.8110112](https://doi.org/10.1109/EWDTS.2017.8110112)
- [10] V. M. Artyushenko, and V. I. Volovach, "Comparative analysis of discriminators efficiency of tracking meters under influence of non-Gaussian broadband and band pass noise", Proceedings XI International IEEE Scientific and Technical Conference "Dynamics of Systems, Mechanisms and Machines (Dynamics)", Nov 14-16, 2017. DOI: [10.1109/Dynamics.2017.8239430](https://doi.org/10.1109/Dynamics.2017.8239430)
- [11] V. M. Artyushenko, and V. I. Volovach, "Measuring information signal parameters under additive non-Gaussian correlated noise", Optoelectronics, Instrumentation and Data Processing, 2016, Vol. 59, No. 6, pp. 22-28. DOI: [10.15372/AUT20160603](https://doi.org/10.15372/AUT20160603)

Planar Butler Matrix Based on Compact Taps

Denis A. Letavin
 Ural Federal University
 Yekaterinburg, Russia
 d.a.letavin@urfu.ru

Abstract—The paper presents a planar Butler matrix operating in the range of 0.9-1.05 GHz, implemented on compact directional couplers. The intersecting lines in the matrix are implemented based on the crossover. Modeling and electrodynamic analysis of the circuits was performed in the NI-AWR Design Environment program. A model of the proposed matrix was also made using photolithography. Experimental and theoretical results have good agreement. The area of the Butler matrix is 59.6% smaller than the area of the standard scheme, with the band reduced by only 15%.

Keywords—coupler, miniaturization, artificial transmission line, compact.

I. INTRODUCTION

Butler matrix is one of the varieties of diagram-forming schemes, and are widely used in antenna technology, to form a fan pattern. Such matrices become attractive when it is necessary to ensure the formation of several beams of the radiation pattern at known phase shifts, since such an implementation does not require phase shifters and any mounting of the printed circuit Board. Planar matrices are an important element of multibeam antenna arrays, and their constructive improvement and improvement of electrical parameters is an urgent task today.

Developers of microwave devices decimeter wavelength range, recently attracted, artificial transmission lines (ATL). First of all, this is due to the possibility of a significant reduction in the size of the structure, without large losses in the performance of the device itself. ATL have a shorter length in relation to the standard segments used in traditional designs of microstrip devices.

The Butler matrix contains $N=2m$ inputs (m is an integer) of the same number of outputs and feeds the antenna array containing N emitters, forming N orthogonal rays. Matrix topologies in the lower part of the decimeter range occupy a large area, for this reason they should be miniaturized. To date, many different Butler matrices with small dimensions can be found in the literature, for example in [1]-[9].

In this paper, instead of traditional directional couplers, compact couplers with similar functionality were used in the matrix scheme. This step will significantly reduce the size of the matrix while maintaining its characteristics within a wide range.

II. DESIGN COMPACT COUPLER

The aim of the work is to simulate a compact Butler matrix with wide range of characteristics. To achieve the goal, the calculation and development of compact couplers was carried out, modeling and optimization of these structures in the NI-AWR environment was carried out. Figure 1 shows a block diagram of the Butler matrix 4x4 (where BLC – directional coupler, FS – phase shifter, Crossover). It consists of 3 dB directional couplers and a crossover. It should be borne in mind that each of these elements introduces its own errors, both amplitude and

phase. The substrate material is FR4 with a dielectric permittivity of 4.4 and a thickness of 1 mm.

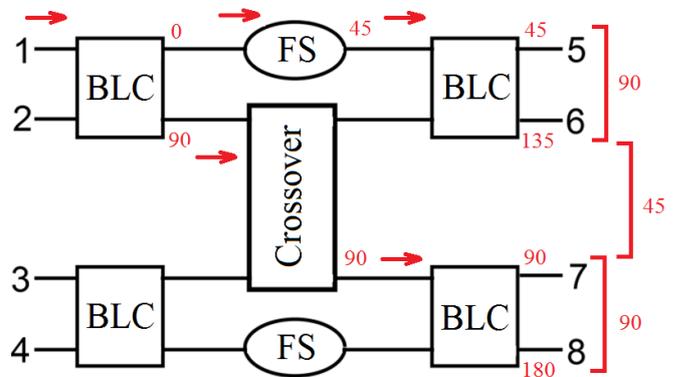


Fig. 1. Block diagram of the 4x4 Butler matrix

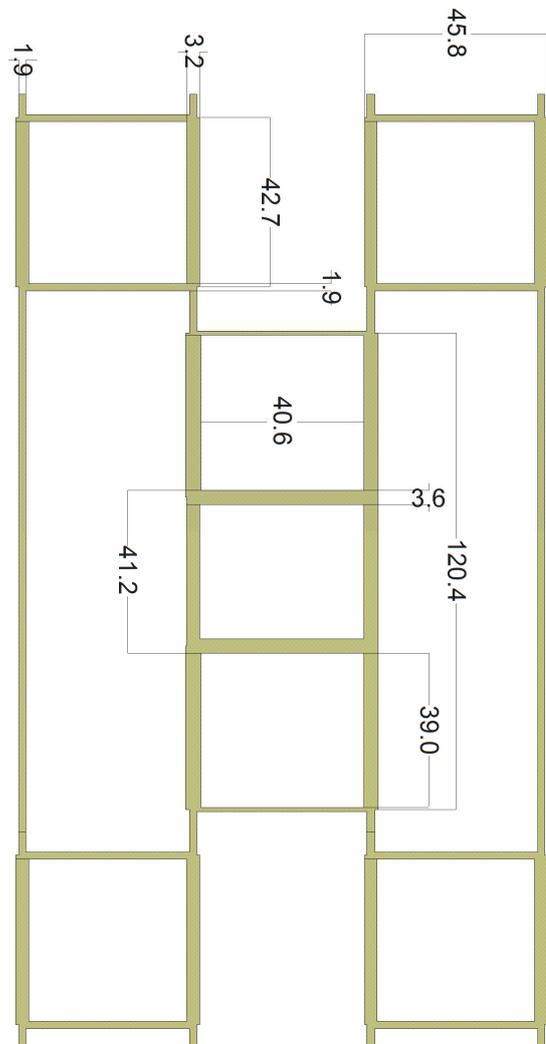


Fig. 2. The topology of the 4x4 Butler matrix for microstrip lines standard

The research was executed by the grant of the Ministry of education and science of the Russian Federation (project № 8.2538.2017/4.6).

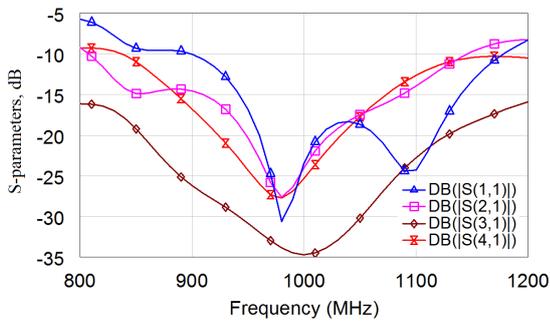


Fig. 3. S-parameters from the frequency

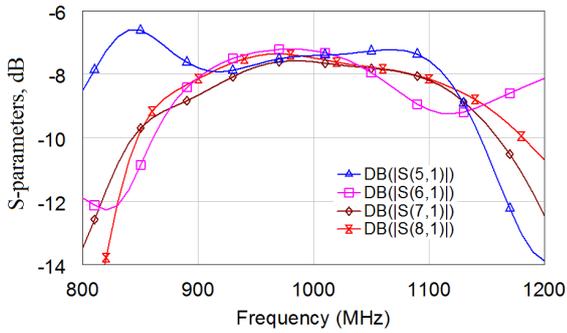


Fig. 4. S-parameters from the frequency

Standard construction occupies an area of $132.7 \text{ mm} \times 230.8 \text{ mm} = 30627.16 \text{ mm}^2$. From the obtained frequency dependences it can be seen that when the first input of the device is excited, its power is divided between its outputs in almost equal proportions near the value -7 dB . The bandwidth determined by the balance of the arms (with an error of $+1 \text{ dB}$) is equal to 100 MHz and is determined primarily by the bands used couplers. In the frequency range from 900 to 1150 MHz , the reflection coefficient and the isolation are below -10 dB .

Based on the fact that the standard topology takes up a lot of space, it was miniaturized through the use of ATL. These lines have similar frequency characteristics, but occupy less space, due to which, with proper arrangement of elements, high results in miniaturization of the matrix are achieved. In the development of compact devices it is necessary to consider the possibility of production of printed circuit boards, and therefore the need to keep technologically feasible gaps and the thickness of the lines. Figure 5 shows a comparison of the phase characteristics of the ATL and the quarter-wave segment of the transmission line. The graph shows that the phases of both lines have a value of 90 degrees. Since the ATL act as a low-pass filter, they also carry out the suppression of higher harmonics.

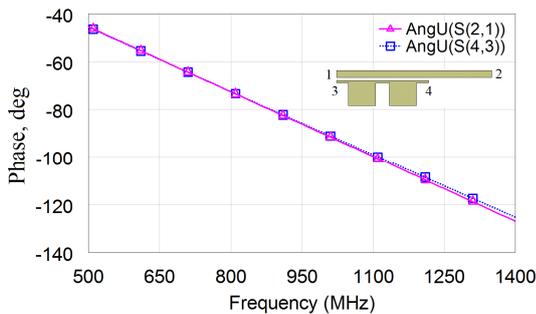


Fig. 5. S-parameters from the frequency

Since the ATL have similar transmission coefficients in the required frequency band with the transmission coefficients of conventional lines, when they are installed, they will not worsen the characteristics of the matrix. The topology of the compact matrix obtained in NI-AWR is illustrated in figure 6, and its frequency characteristics in figures 7,8.

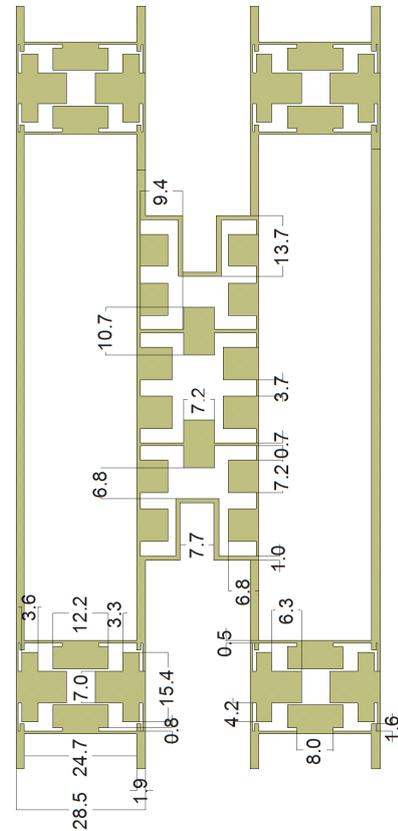


Fig. 6. The topology of compact 4x4 Butler matrix for ATL

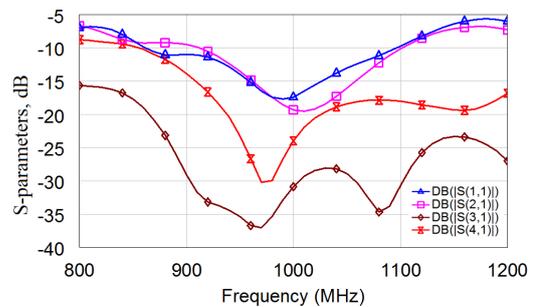


Fig. 7. S-parameters from the frequency

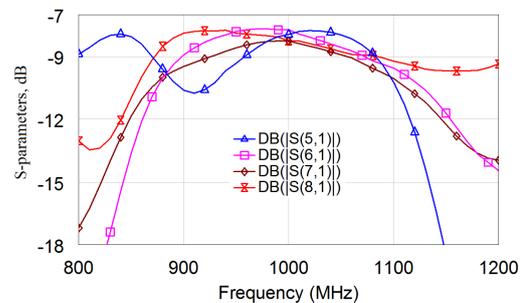


Fig. 8. S-parameters from the frequency

The area of the compact Butler matrix is $80.3 \text{ mm} \times 154.1 \text{ mm} = 12374.23 \text{ mm}^2$, which is 59.6% less than the area of the standard matrix. The frequency characteristics obtained as a result of the electrodynamic analysis of the circuit showed that when the first input of the device is excited, its power is divided between its outputs in almost equal proportions near the value -8 dB . The bandwidth is 80 MHz. In the frequency band from 900 to 1100 MHz, the reflection coefficient and the isolation are below -10 dB . It can be seen that there was a decrease in transmission coefficients and narrowing of the frequency band, these are negative miniaturization factors.

III. PROTOTYPE MEASUREMENTS

After the topology of the compact matrix is obtained, with the help of photolithography we make a layout of the scheme to check its performance in practice. Figure 9 shows a photograph of the manufactured circuit. The experimental characteristics obtained with the help of the vector network analyzer Rohde & Schwarz ZVA24 are shown in figures 10, 11.

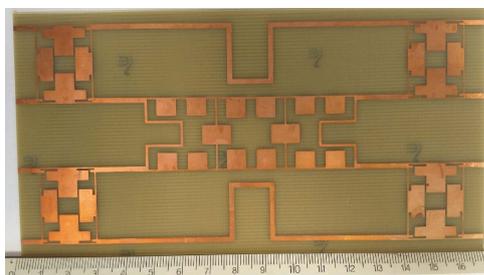


Fig. 9. Photo of the prototype of a compact coupler

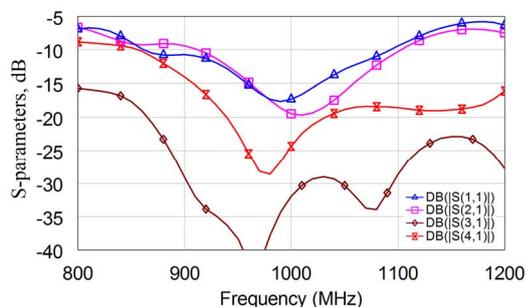


Fig. 10. The measured S-parameters of a compact coupler

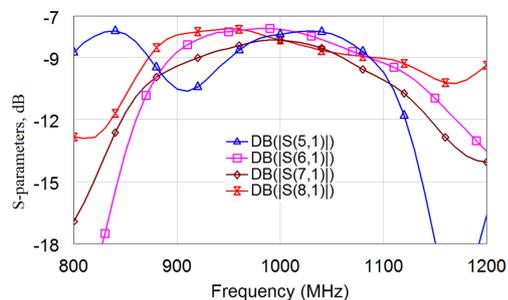


Fig. 11. The measured S-parameters of a compact coupler

From the obtained frequency dependences it can be seen that when the first input of the device is excited, its power is divided between its outputs in almost equal proportions about the value -8.5 dB . The bandwidth is 85 MHz. In the frequency band from 900 to 1080 MHz reflection and decoupling coefficients are below -10 dB . When comparing

the experimental and theoretical characteristics can be seen that they have a high similarity. Also, the main results are summarized in table 1.

TABLE I. COMPARISON OF DESIGN STRUCTURES

Design	Bandwidth, MHz	Area, mm ²	Size reduction, %	Output phases, degree
Standard	100	30627.2	-	90,45,90
Compact	85	12374.2	59.6	88,46,88

IV. CONCLUSION

As a result, a compact planar matrix of the Butler 4x4 was investigated and developed. It can be used to power antenna arrays and form a fan pattern. Due to the use of compact couplers, it was possible to reduce the total area of the scheme by 59.6% in relation to the area of the standard matrix. The model of the proposed matrix was made, and its frequency characteristics were measured, which showed good compliance with the theoretical modeling. The disadvantages of miniaturization is an increase in losses by 1 dB and a narrowing of the frequency band by 15%. The proposed matrix has compact dimensions that can be provided during mass production.

ACKNOWLEDGMENT

Author are grateful for the simulation software NIAWR Design Environment provided by the National Instruments Company.

REFERENCES

- [1] J. Butler, R. Lowe, "Beam-forming matrix simplifies design of electronically scanned antennas," *Electron. Des.*, vol. 9, pp. 170-173, April 1961.
- [2] K. Winca, S. Gruszczynski, "A broadband 4 x 4 Butler matrix for modern-day antennas," in *Proc. 35th European Microwave Conference*, Paris, France, Oct. 2005, pp. 1331-1334.
- [3] S. Gruszczynski, K. Winca, K. Sachse, "Compact broadband Butler matrix in multilayer technology for integrated multibeam antennas," *Electronics Letters*, vol. 43, no. 11, pp. 635-636, May 2007.
- [4] George Tudose, Helmut Barth, Rudiger Vahldieck, "A Compact LTCC Butler Matrix Realization for Phased Array Applications," 2006 IEEE MTT-S International Microwave Symposium Digest, DOI: 10.1109/MWSYM.2006.249586.
- [5] Changfei Zhou, Jiahui Fu, Haifeng Sun, Qun Wu, "A novel compact dual-band butler matrix design," *Proceedings of 2014 3rd Asia-Pacific Conference on Antennas and Propagation*, DOI: 10.1109/APCAP.2014.6992767
- [6] Han Ren, Jin Shao, Rongguo Zhou, Bayaner Arigong, Hyoung Soo Kim, Changzhi Li, Hualiang Zhang, "A compact phased array antenna system based on dual-band Butler matrices," *Proceedings of 2014 3rd Asia-Pacific Conference on Antennas and Propagation*, DOI: 10.1109/RWS.2013.6486692
- [7] P.Q. Mariadoss, M.K.A. Rahim, M.Z.A.A. Aziz, "Design and implementation of a compact Butler matrix using mitered bends," 2005 Asia-Pacific Microwave Conference Proceedings, DOI: 10.1109/APMC.2005.1606980
- [8] J. J. Kuek, Karthik T. Chandrasekaran, M. F. Karim, Nasimuddin, A. Alphones, "A compact Butler matrix for wireless power transfer to aid electromagnetic energy harvesting for sensors," 2017 IEEE Asia Pacific Microwave Conference (APMC), DOI: 10.1109/APMC.2017.8251447
- [9] Letavin, D.A., Mitelman, Y.E., Chechetkin, V.A., "Compact microstrip branch-line coupler with unequal power division", 2016 24th Telecommunications Forum (TELFOR), 2016, TELFOR 2016, DOI:10.1109/TELFOR.2016.7818850

All-Pass Second-Order Active RC-Filter with Pole Q-Factor's Independent Adjustment on Differential Difference Amplifiers

Darya Yu. Denisenko
Don State Technical University,
Southern Federal University,
Rostov-on-Don, Taganrog, Russia
d.y.denisenko@yandex.ru

Nikolay N. Prokopenko
Member, IEEE
Don State Technical University,
IPPM RAS,
Rostov-on-Don, Zelenograd, Russia
prokopenko@sssu.ru

Nikolay V. Butyrlagin
Don State Technical University
Rostov-on-Don, Russia
butyrlagin@gmail.com

Abstract— The article presents a scheme of active RC-filter of the second order (ARCF), which is realized on multidifferential operational amplifiers (DDA). Based on this filter, we can obtain a various amplitude-frequency characteristic (AFC) (low-pass filters (LPF) and high-pass (HPF) filters, band-pass (BPF) and rejection (RF) filters). The main difference between the developed ARCF and the known ones is that when adjusting the quality factor of the pole, their pole frequency and transmission coefficient do not change.

Keywords—Active Filters, Low-Pass Filters, High-Pass Filters, Band-Pass Filters, Reject Filters, All-Pass Filters, Anti-Alias Filters, Analog Digital Converters, Wide-Band Selective Amplifiers, Differential Difference Amplifiers, Pole Q-Factor, Pole Frequency, Transfer Ratio, Parameters' Trimming

I. INTRODUCTION

For the tasks of isolating and limiting the spectrum of signals at the input of analog-to-digital converters [1,2] it is possible to use universal active RC-filters (ARCF) [3-9], which allow to realize different amplitude-frequency characteristics (AFC) of low-pass filter, high-pass filter, PF, RF. At the same time, the application of new electronic component base in ARCF, for example, on differential difference operational amplifiers (DDA) [10-15], providing new qualities to frequency selection devices, is of considerable interest.

The purpose of this work is to study the new structure of the universal ARCF [16], which provides independent regulation of the main parameters (for example, by means of digital switching of passive elements or digital potentiometers).

II. ARC-FILTER'S BASE ARCHITECTURE'S PROPERTIES

On Fig. 1 shows the structure of the proposed universal ARC-filter implementing LPF, HPF, BPF, RF [16]. The main feature of the scheme Fig.1 is an independent adjustment of the quality factor of the pole, and such filter parameters as the frequency of the pole and the transmission coefficient remain unchanged. These advantages make it possible to simplify of the tuning process and tweaking the basic parameters of frequency selection devices.

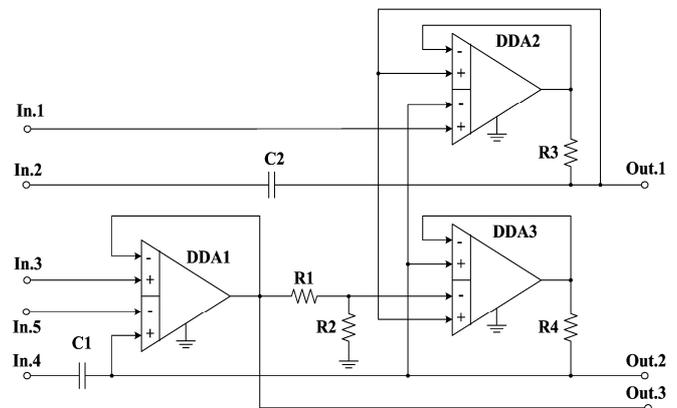


Fig. 1. All-pass ARC-Filter's Circuit.

The transfer function for different types of ARCF (LPF, HPF, RF, BPF) implemented at using the circuit (Fig. 1), is presented below

$$F(p) = \frac{a_2 p^2 + a_1 p + a_0}{b_2 p^2 + b_1 p + b_0}, \quad (1)$$

where a_i, b_j are numerator's (Table 1) and denominator's (2) coefficients.

Using a set of coefficients a_i of the numerator of the transfer function is determined by the types of ARCF (LPF, HPF, BPF, RF). The coefficients of the denominators of the transfer functions b_j (1) of the diagram Fig. 1 described by these formulas

$$b_2 = 1, \quad b_1 = \frac{K}{\tau_1}, \quad b_0 = \frac{1}{\tau_1 \tau_2}, \quad (2)$$

where $K = \frac{R_2}{R_1 + R_2}$, $\tau_1 = R_4 C_1$, $\tau_2 = R_3 C_2$.

The research is made within the Russian Science Foundation Grant (Project No. 18-79-10109)

TABLE I. ARC-FILTERS' TRANSFER FUNCTIONS' (I) NUMERATOR COEFFICIENTS A_i, IMPLEMENTED ON CIRCUIT FIG. 1

		OUTPUTS		
		1	2	3
INPUTS	1	$\text{BPF}^{(+)}+\text{LPF}^{(-)}$ $a_2 = 0$ $a_1 = \frac{1}{\tau_2}$ $a_0 = \frac{K}{\tau_1\tau_2}$	$\text{LPF}^{(+)}$ $a_2 = a_1 = 0$ $a_0 = \frac{1}{\tau_1\tau_2}$	$\text{LPF}^{(+)}$ $a_2 = a_1 = 0$ $a_0 = \frac{1}{\tau_1\tau_2}$
	2	$\text{HPF}^{(+)}+\text{BPF}^{(-)}$ $a_2 = 1$ $a_1 = \frac{K}{\tau_1}$ $a_0 = 0$	BPF $a_2 = a_1 = 0$ $a_1 = \frac{1}{\tau_1}$	BPF $a_2 = a_1 = 0$ $a_1 = \frac{1}{\tau_1}$
	3	LPF $a_2 = a_1 = 0$ $a_0 = \frac{K}{\tau_1\tau_2}$	$\text{BPF}^{(+)}$ $a_2 = a_0 = 0$ $a_1 = -\frac{K}{\tau_1}$	$\text{RF}^{(+)}$ $a_2 = 1$ $a_1 = 0$ $a_0 = \frac{1}{\tau_1\tau_2}$
	4	BPF $a_2 = a_0 = 0$ $a_1 = -\frac{1}{\tau_2}$	$\text{HPF}^{(+)}$ $a_2 = 1$ $a_1 = a_0 = 0$	$\text{HPF}^{(+)}$ $a_2 = 1$ $a_1 = a_0 = 0$
	5	LPF $a_2 = a_1 = 0$ $a_0 = -\frac{K}{\tau_1\tau_2}$	$\text{BPF}^{(+)}$ $a_2 = a_0 = 0$ $a_1 = \frac{K}{\tau_2}$	$\text{RF}^{(+)}$ $a_2 = -1$ $a_1 = 0$ $a_0 = -\frac{1}{\tau_1\tau_2}$
	1 & 5	LPF $a_2 = a_1 = 0$ $a_0 = \frac{K}{\tau_1\tau_2}$	$\text{RF}^{(+)}$ $a_2 = 1$ $a_1 = 0$ $a_0 = \frac{1}{\tau_1\tau_2}$	$\text{RF}^{(+)}$ $a_2 = 1$ $a_1 = 0$ $a_0 = \frac{1}{\tau_1\tau_2}$
	4 & 5	$\text{BPF}^{(+)}+\text{LPF}^{(-)}$ $a_2 = 0$ $a_1 = -\frac{1}{\tau_2}$ $a_0 = -\frac{K}{\tau_1\tau_2}$	$\text{HPF}^{(+)}+\text{BPF}^{(-)}$ $a_2 = 1$ $a_1 = \frac{K}{\tau_1}$ $a_0 = 0$	$\text{LPF}^{(+)}$ $a_2 = a_1 = 0$ $a_0 = -\frac{1}{\tau_1\tau_2}$

The active RC-filters presented in Table 1 ($LPF^{(+)}$, $HPF^{(+)}$, $BPF^{(+)}$, $RF^{(+)}$) provide independent tuning of the main filter parameters - transmission coefficient, frequency and pole quality). Moreover, when adjusting the quality factor of the filter pole, the transmission coefficient and the pole frequency do not change [16]. In practical terms, these features can be widely used.

The Table 1 presents the active RC-filters (LPF, HPF, BPF, RF) that do not have the properties of independent tuning of the main filter parameters (pole quality, transmission coefficient and pole frequency) [16]. When configuring these active filters (as well as classical active filters of the second order), when one of the parameters changes, for example, the pole quality factor, other parameters (transfer coefficient and pole frequency) can change.

The Table 1 also contains active RC-filters ($LPF^{(-)}$, $HPF^{(-)}$, $BPF^{(-)}$, $RF^{(-)}$), which have a small slope of the amplitude-frequency response, which corresponds to the transfer function of the first order [16] and is a disadvantage of these schemes.

In fig. 2 presents various options for enabling the inputs and outputs of the ARCF (Fig. 1), allowing to obtain various modifications of the frequency response. Moreover, unused inputs should be connected to the common power supply bus.

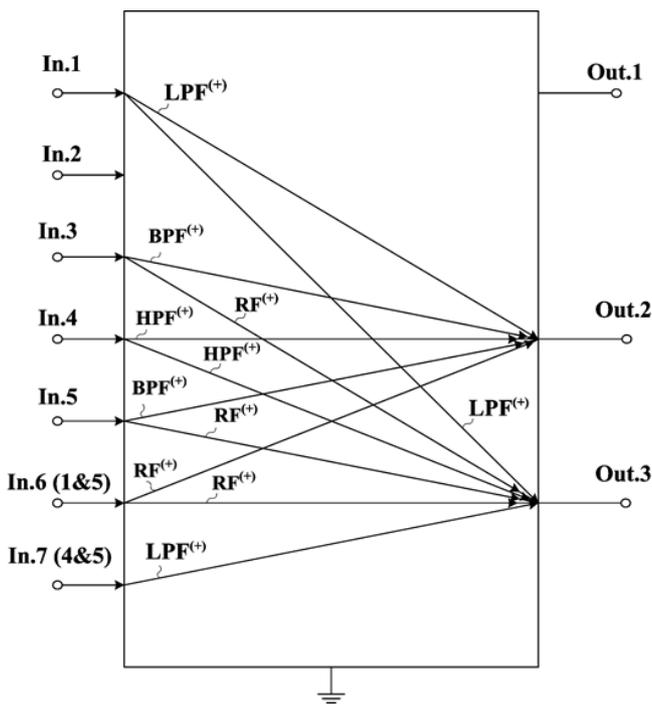


Fig. 2. ARCF Fig.1 Inputs and Outputs Application Variants, which Provide $BPF^{(+)}$, $RF^{(+)}$, $LPF^{(+)}$, $HPF^{(+)}$ Implementation with Pole Q-factor Independent Tuning.

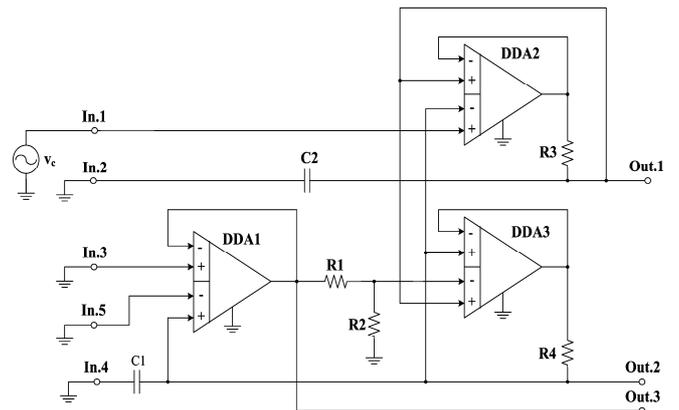
III. EXAMPLE OF CREATING LOW-PASS FILTER

In this section, we consider two examples of the implementation of a low-pass filter [16] based on DDA. The AFCs of active RC-filters ($PF^{(-)} + LPF^{(-)}$, $LPF^{(+)}$, $LPF^{(+)}$), based in the circuit of Fig. 3a for outputs Out.1, Out.2 and Out.3, respectively, are shown in Fig. 3b.

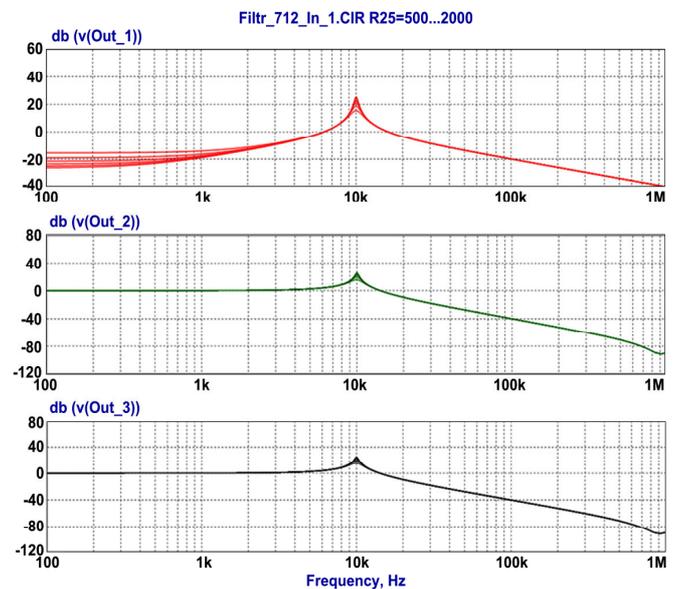
Analysis of AFCs graphs Fig. 3b shows that for outputs Out.2 and Out.3 when adjusting the pole Q-factor, the transfer ratio and pole frequency do not change.

In Fig. 4 shows the second ARC-filter enable Fig. 1 and its AFCs for ARC-filters ($PF^{(-)} + LPF^{(-)}$, $HPF^{(-)} + PF^{(-)}$, $LPF^{(+)}$), implemented in the scheme of Fig. 4a for Out.1, Out.2 and Out.3 respectively.

As well as in the scheme of Fig. 3a, analysis of AFCs graphs Fig. 4b showed that in the Out3 $LPF^{(+)}$ when adjusting the pole Q-factor, the transfer ratio and pole frequency do not change.



(a)



(b)

Fig. 3. The First Special Case of the Inclusion of the ARC-Filter Fig. 1 (a) and its AFCs (b).

The computer simulation in the MicroCap environment on models of DDA AD830 for other ARCF modifications (HPF, PF, RF) [16], which are implemented on the basis of the scheme Fig. 1, confirmed the above distinctive properties of the proposed circuit solutions.

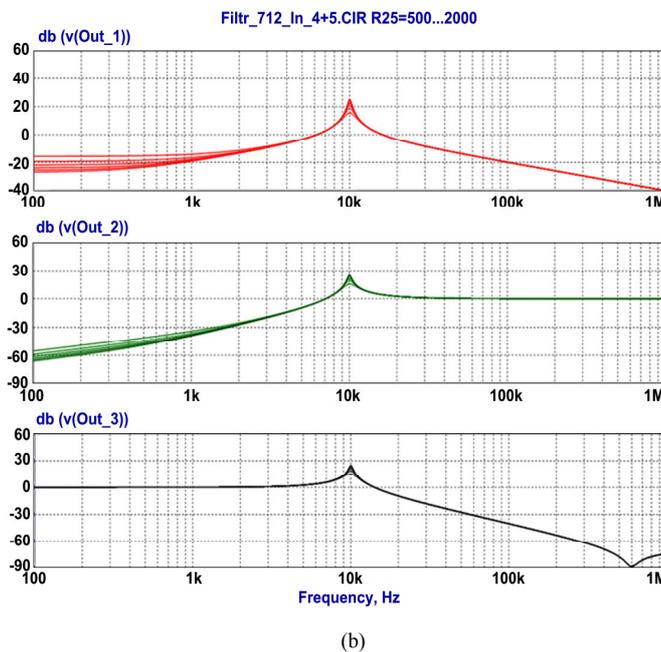
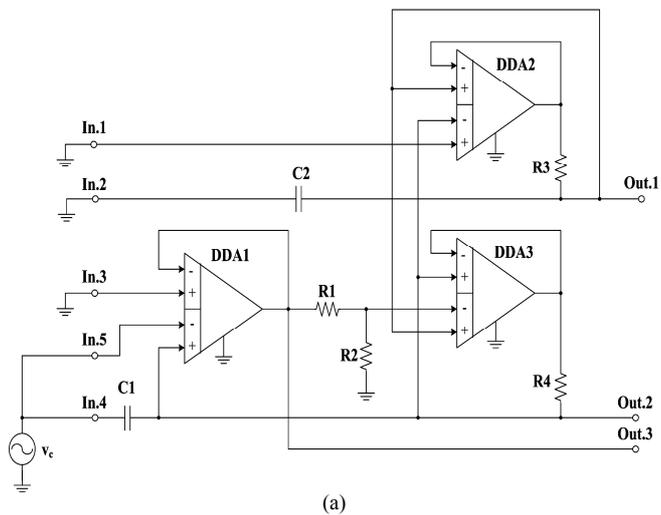


Fig. 4. The Second Special Case of the Inclusion of the ARC-Filter Fig. 1 (a) and its AFCs (b).

IV. CONCLUSION

The structure of the universal ARCF, which allowing to implement various second-order filters (LPF, HPF, BPF, RF), is presented. Moreover, a distinctive feature of these filters is the possibility of independent adjustment of the main parameters (transmission coefficient, quality factor and pole frequency). It is a great advantage of the suggested schematic solution, which is recommended to implement specialized base structural crystal in a form of circuits within tasks of frequency selection and signals' spectrum limitation on modern analog digital converters' inputs.

REFERENCES

[1] I. V. Vernik et al., "Superconducting High-Resolution Low-Pass Analog-to-Digital Converters," in *IEEE Transactions on Applied Superconductivity*, vol. 17, no. 2, pp. 442-445, June 2007. DOI: 10.1109/TASC.2007.898613

[2] M. A. E. Latina, M. P. Sejera, J. P. V. Mitra, B. S. Monton and C. S. Pundan, "A Study of the Effect of Integrating Low-Pass Filter in Measuring the Dynamic Performance of a High Speed 8-bit Analog-to-Digital Converter (ADC)," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 2018, pp. 1-6. DOI: 10.1109/HNICEM.2018.8666282

[3] H. Tarunkumar, A. Ranjan and N. M. Pheiroijam, "Fourth Order Band Pass and All Pass Filter using Single FTFN," 2018 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2018, pp. 1-4. DOI: 10.1109/ICCCI.2018.8441281; WOS:000447639800034

[4] A. Paul, J. Ramírez-Angulo, A. J. Lopez-Martin and R. Gonzalez Carvajal, "CMOS First-Order All-Pass Filter With 2-Hz Pole Frequency," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 2, pp. 294-303, Feb. 2019. DOI: 10.1109/TVLSI.2018.2878017; WOS:000458069300003

[5] K. Roja, M. Sarada and A. Srinivasulu, "A Voltage Mode All Pass Filter Employing Voltage Difference Transconductance Amplifier," 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Iasi, Romania, 2018, pp. 1-4. DOI: 10.1109/ECAI.2018.8679013

[6] S. Lin, X. Zuo and X. Deng, "Current and Voltage Mode Resistorless Universal Biquad Filter Using a Single CCCDTA," in *Chinese Journal of Electronics*, vol. 27, no. 6, pp. 1250-1257, 11 2018. DOI: 10.1049/cje.2018.08.007; WOS:000451790000020

[7] W. Cheta, M. Siripruchyanun, K. Trachu, P. Suwanjan, R. Sotner and W. Jaikla, "Single VDCC Based Voltage-mode First-order Allpass Filter with Electronic Controllability," 2018 18th International Symposium on Communications and Information Technologies (ISCIT), Bangkok, 2018, pp. 255-260. DOI: 10.1109/ISCIT.2018.8587939; WOS:000457691400051

[8] A. Asoodeh and S. Mirabbasi, "On the Design of n th-Order Polyphase All-Pass Filters," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 1, pp. 133-146, Jan. 2019. DOI: 10.1109/TCSI.2018.2853632

[9] B. Singh, A. Kumar Singh, R. Senani, "A new universal biquad filter using differential difference amplifiers and its practical realization," *Analog Integr Circ Sig Process* (2013) 75: 293. DOI: 10.1007/s10470-013-0048-4; WOS:000317623500012

[10] R. Sotner, N. Herencsar, V. Kledrowetz, A. Kartci and J. Jerabek, "New Low-Voltage CMOS Differential Difference Amplifier (DDA) and an Application Example," 2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS), Windsor, ON, Canada, 2018, pp. 133-136. DOI: 10.1109/MWSCAS.2018.8623866

[11] C. Bauzá, J. M. Sánchez-Chiva, J. Madrenas and D. Fernández, "Optimizing Power Consumption vs. Linearization in CMFB Amplifiers with Source Degeneration," 2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Bordeaux, 2018, pp. 269-272. DOI: 10.1109/ICECS.2018.8617958

[12] B. A. Minch, "A CMOS differential-difference amplifier with class-AB input stages featuring wide differential-mode input range," 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, 2017, pp. 1-4. DOI: 10.1109/ISCAS.2017.8050488

[13] J. S. Mincey, C. Briseno-Vidrios, J. Silva-Martinez and C. T. Rodenbeck, "Low-Power Gm-C Filter Employing Current-Reuse Differential Difference Amplifiers," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 6, pp. 635-639, June 2017. DOI: 10.1109/TCSII.2016.2599027

[14] A. J. Lopez-Martin, M. P. Garde and J. Ramirez-Angulo, "Class AB differential difference amplifier for enhanced common-mode feedback," in *Electronics Letters*, vol. 53, no. 7, pp. 454-456, 30 3 2017. DOI: 10.1049/el.2017.0347

[15] F. Khateb and T. Kulej, "Design and Implementation of a 0.3-V Differential Difference Amplifier," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 2, pp. 513-523, Feb. 2019. DOI: 10.1109/TCSI.2018.2866179; WOS:000457357200006

[16] D. Yu. Denisenko, N. N. Prokopenko, "All-Pass Active RC-Filter," Patent appl. RU 2019107174, Filed: March 14, 2019.

Planar Compact Directional Coupler on Artificial Transmission Lines

Denis A. Letavin
 Ural Federal University
 Yekaterinburg, Russia
 d.a.letavin@urfu.ru

Abstract— A microstrip directional coupler with reduced physical dimensions while maintaining the frequency characteristics within a wide range was developed. The area of the compact coupler is 75.3% less than the area of the traditional one. In a design that can be used in telecommunications, the standard quarter-wave segments are replaced by artificial transmission lines. The entire modeling process was performed in the NI-AWR Design Environment program. With the help of photolithography, a model of the proposed device was made. It is also shown that the results obtained during field experiments are in good agreement with the theoretical results.

Keywords—coupler, miniaturization, artificial transmission line, compact.

I. INTRODUCTION

Developers of microwave devices decimeter wavelength range, recently attracted artificial transmission lines (ATL). First of all, this is due to the possibility of a significant reduction in the size of the structure, without large losses in the performance of the device itself. And ATL have a shorter length in relation to the standard segments used in the design of the taps. When the operating frequency decreases, the length of the quarter-wave segments increases, and this leads to an increase in the area that occupies the device. To date, many methods have been developed to reduce the size of microstrip devices [1]-[9]. This paper presents the design of the developed compact microstrip directional coupler, in which all quarter-wave segments are replaced by equivalent ATL. The whole process of modeling the coupler was carried out using a specialized software product NI-AWR Design Environment.

II. DESIGN COMPACT COUPLER

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file. The aim of the work is to develop a compact directional coupler decimeter wavelength range. To achieve the goal, the calculation and development of the coupler was carried out, modeling and optimization of the structure in the NI-AWR environment was carried out. Figure 1 shows the flow diagram of the directional coupler.

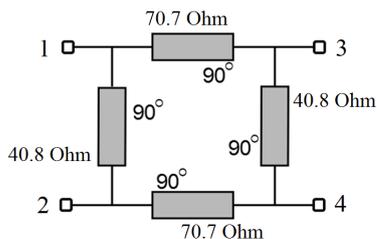


Fig. 1. A block diagram of the directional coupler

Initially, with the help of AVR, a standard design of a directional coupler was designed. Wave resistances of this design were calculated using the following formulas [10]:

$$\begin{aligned} Z_1 &= Z_{in} k, \\ Z_2 &= Z_{in} \frac{k}{\sqrt{1-k^2}}, \end{aligned} \quad (1)$$

where Z_{in} – input impedance of all ports,

$$\frac{k^2}{1-k^2} = \frac{P_3}{P_4} \quad - \text{ power ratio between the outputs,}$$

$k = \sqrt{\frac{P_4}{P_3}}$. So if $Z_{in} = 50 \text{ Ohm}$, $P_3 / P_4 = 2$, then $k = \sqrt{2/3}$, $Z_1 = 40.8 \text{ Ohm}$, $Z_2 = 70.7 \text{ Ohm}$.

The topology of the deflector on standard microstrip lines, with a Central frequency of 1 GHz, is shown in figure 2. The substrate is FR4 material with dielectric permittivity $\epsilon = 4.4$ and thickness $h = 1 \text{ mm}$. the Frequency characteristics for this topology obtained in the AWR program are shown in figures 3,4.

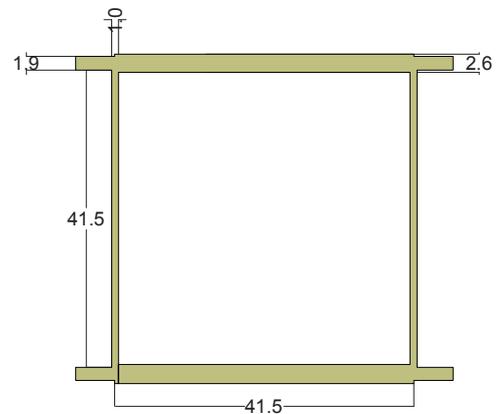


Fig. 2. Topology of the standard directional coupler

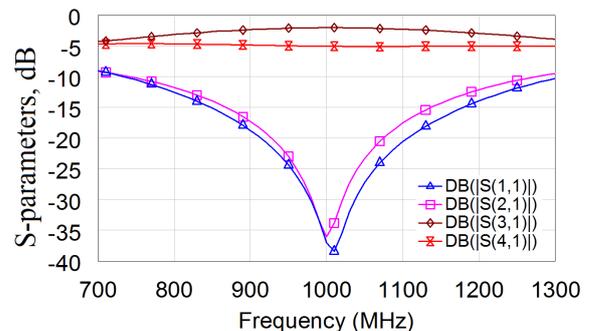


Fig. 3. S-parameters from the frequency

The research was executed by the grant of the Ministry of education and science of the Russian Federation (project № 8.2538.2017/4.6).

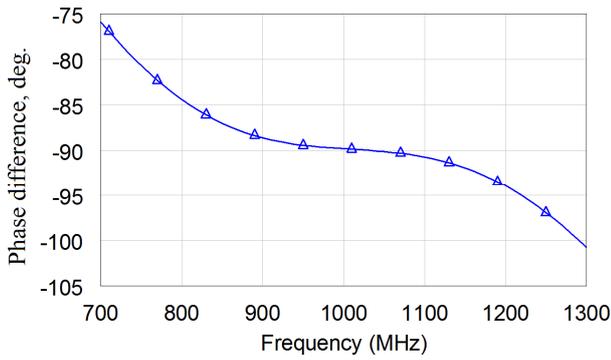


Fig. 4. The phase difference between the transmission coefficients of the frequency

Standard construction covers an area equal to $40.6 \text{ mm} \times 42.5 \text{ mm} = 1725.5 \text{ mm}^2$. From the obtained frequency dependences it can be seen that when the first input of the device is excited, its power is divided between 3 and 4 outputs 1 to 2, and have values of 2.07 and 5.08 dB. The bandwidth level of -20 dB is 147 MHz. The reflection coefficient at the Central frequency of 1 GHz has a value of 37 dB.

Due to the fact that the standard design takes up a lot of space, its dimensions have been miniaturized by the use of ATL. Since the ATL can fit tightly to each other, completely filling the space inside the branch. However, with this arrangement of ATL, there are small distances between adjacent conductive elements, as a result of which some elements have an electromagnetic effect on other elements. It is also worth considering that the gaps between the elements should be technologically simple to implement.

Due to the fact that the ATL have similar characteristics with low-pass filters, they pass the signal in the required band and suppress it at higher frequencies. The proposed ATL is easy to implement. With the use of modern technical means, it is possible to achieve high accuracy of calculation, and with the improvement of photolithography technology will achieve the desired characteristics of the ATL in their manufacture. Figure 5 shows the topology of one of the ATL with its frequency characteristics.

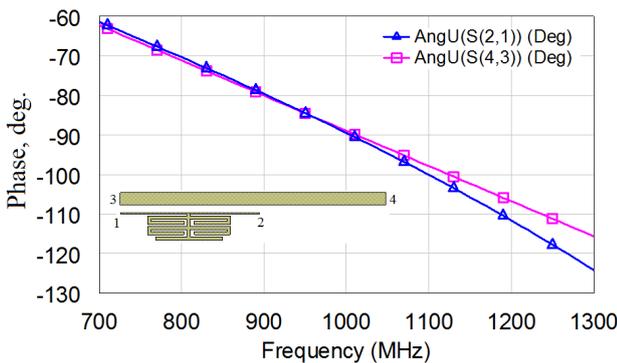


Fig. 5. Chart comparing the phase of the microstrip line and its equivalent line

It can be seen that at the Central frequency, the ATL and the quarter-wave segment have the same phase equal to 90 degrees. Due to this, the ATL can be installed instead of quarter-wave segments without loss of efficiency. The topology of the coupler with installed ATL is shown in figure 6, and its frequency characteristics in figures 7.8.

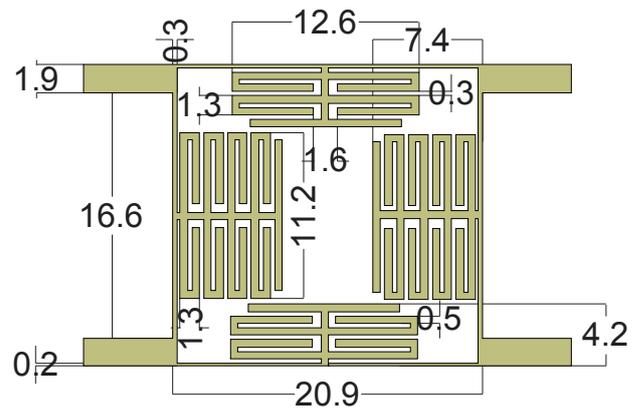


Fig. 6. Topology of the compact directional coupler

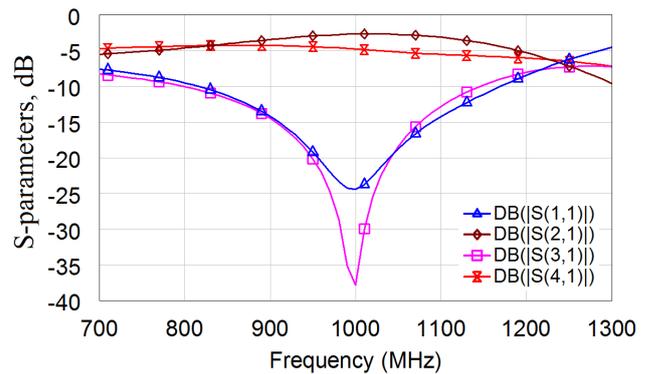


Fig. 7. S-parameters from the frequency

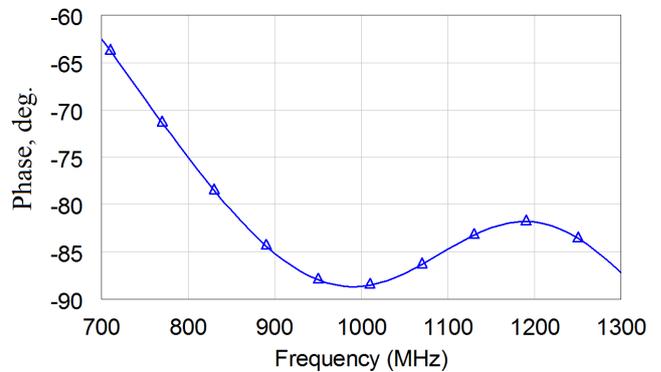


Fig. 8. The phase difference between the transmission coefficients of the frequency

The compact design covers an area of $20.4 \text{ mm} \times 20.9 \text{ mm} = 426.36 \text{ mm}^2$, which is 75.3% less than the area of the traditional design. From the obtained frequency dependences it can be seen that when the first input of the device is excited, its power is divided between the outputs from the values of 2.6 and 4.8 dB. The bandwidth at the level of -20 dB is 91 MHz – this is 38% less than that of the coupler on the microstrip segments. The reflectance at the Central frequency of 1 GHz is 24 dB.

III. PROTOTYPE MEASUREMENTS

To verify the results obtained in the AWR program, the resulting topology was made by milling. Fig.9 shows a photo of the resulting layout. Measurements of electrical parameters of the prototype were carried out in the frequency band from 1 to 3 GHz on the vector network analyzer Rohde & Shwarz ZVA24 (Fig.10,11).

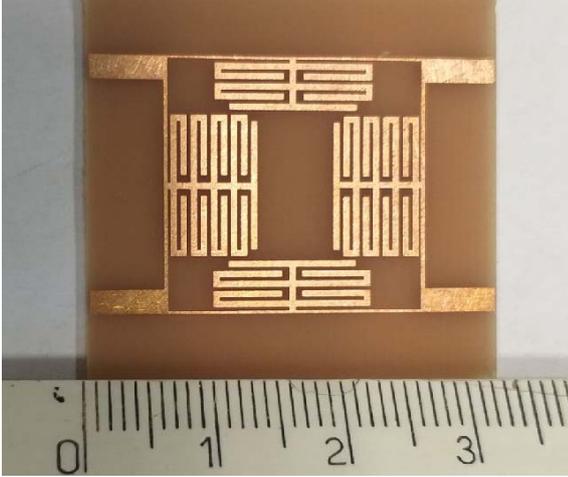


Fig. 9. Photo of the prototype of a compact coupler

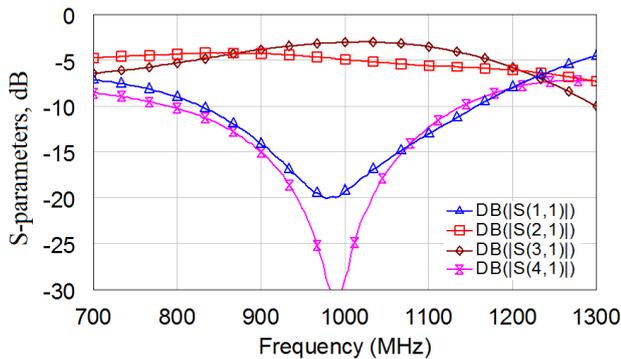


Fig. 10. The measured S-parameters of a compact coupler

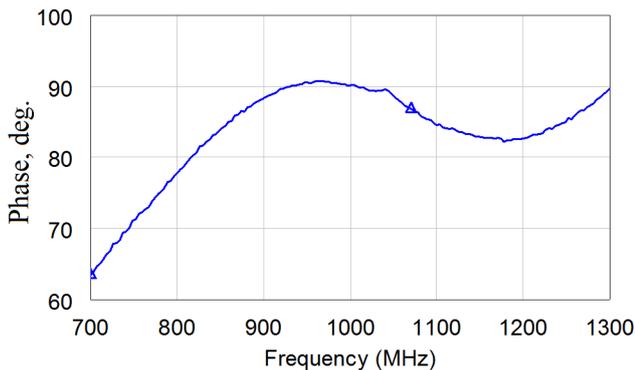


Fig. 11. The measured phase difference between the outputs of a compact coupler

From the obtained frequency dependences it can be seen that when the first input of the device is excited, its power is divided between 2 and 3 with values of 3 and 4.9 dB. The bandwidth at -20 dB is 90 MHz. The reflectance at the Central frequency of 1 GHz is 20 dB. Comparative analysis of the results obtained by electrodynamic calculation and full-scale experiments of the device showed that the results

are consistent. Table 1 summarizes all the simulation and experimental results described.

TABLE I. COMPARISON OF STANDARD AND MINIATURE STRUCTURES

Design	Decoupling band at -20 dB, MHz	Area, mm ²	Size reduction, %
Standard	147	1725.5	-
Compact model	91	426.4	75.3
Compact layout	90	426.4	75.3

IV. CONCLUSION

In this paper we propose a planar topology of a compact coupler with a Central frequency of 1 GHz. The area of this device is reduced by the use of ATL instead of quarter-wave segments of the transmission line. The modeling process was carried out in the program of the NI AWR Design Environment Made the experimental sample and measured its electrical characteristics. The theory on which the synthesis procedure is based is confirmed by the results of modeling and experiment. The area of the compact coupler is 75.3% smaller than the standard one, but its frequency band has been reduced by 38.7%.

ACKNOWLEDGMENT

Author are grateful for the simulation software NIAWR Design Environment provided by the National Instruments Company.

REFERENCES

- [1] S.-S. Liao and J.-T. Peng, "Compact planar microstrip branch-line couplers using the quasi-lumped elements approach with nonsymmetrical and symmetrical T-shaped structure," *IEEE Trans. Microw. TheoryTech.*, vol. 54, pp. 3508-3514, Sep. 2006.
- [2] S.-S. Liao, P.-T. Sun, N.-C. Chin and J.-T. Peng, "A novel compact-size branch-line coupler," *IEEE Microw. Wireless Compon. Lett.* vol. 15, pp. 588-590, Sep. 2005.
- [3] Chao-Wei Wang, Tzyh-Ghuang Ma and Chang-Fa Yang, "A new planar artificial transmission line and its applications to a miniaturized butler matrix," *IEEE Trans. Microw. Theory Tech.*, vol. 55, no. 12, pp. 2792-2801, Dec. 2007.
- [4] H. Ghali, and T. A. Moselhy, "Miniaturized fractal ratrace, branch-line, and coupler-line hybrids," *IEEE Trans. Microw. Theory Tech.*, vol. 52, pp. 2513-2520, Nov. 2004.
- [5] Chao-Hsiung Tseng and Chih-Lin Chang, "A rigorous design methodology for compact planar branch-line and rat-race couplers with asymmetrical T-structure," *IEEE Trans. Microw. Theory Tech.*, vol. 60, no. 7, pp. 2085-2092, July 2012.
- [6] Kai-Yu Tsai, Hao-Shun Yang, Jau-Horng Chen, and Yi-Jan Emery Chen, "A miniaturized 2 dB Branch-Line Hybrid Coupler With Harmonics Suppression," *IEEE Microw. Wireless Compon. Lett.*, vol. 21, no. 10, pp. 537-539, Oct. 2011.
- [7] Ashmi Chakraborty Das, Lakhindar Murmu, Santau Dwari, "A Compact Branch-Line Coupler Using Folded Microstrip Lines," *IEEE Microw. Wireless Compon. Lett.*, vol. 10, no. 7, pp. 1-3, Dec. 2013.
- [8] Hani Ghali and Tarek A. Moselhy "Miniaturized fractal rat-race, branch-line, and coupled-line hybrids," *IEEE Trans. Microw. Theory Tech.*, vol. 52, no. 11, pp. 2513-2520, Nov. 2004.
- [9] Letavin, D.A., Mitelman, Y.E., Chechetkin, V.A., "Investigation of the frequency influence on the miniaturization efficiency of microstrip devices using LPFs", 2016 10th European Conference on Antennas and Propagation, EuCAP 2016, DOI: 10.1109/EuCAP.2016.7481614.
- [10] C. Tokar; M. Saglam; M. Ozme; N. Gunalp, "Branch-line couplers using unequal line lengths", *IEEE Transactions on Microwave Theory and Techniques*, 2001, Vol. 49, pp. 718-721.

Silicon Photomultipliers' Analog Interface with Wide Dynamic Range

Oleg V. Dvornikov
Minsk Research Instrument-Making
Institute JSC (MNIPI JSC),
Minsk, Belarus
oleg_dvornikov@tut.by

Yaroslav D. Galkin
Belarusian State University of
Informatics and Radioelectronics,
Minsk, Belarus
galkinyaroslav@gmail.com

Nikolay N. Prokopenko
Member, IEEE
Don State Technical University, IDPM
Institute for Design Problems in
Microelectronics of RAS,
Rostov-on-Don, Zelenograd, Russia
prokopenko@sssu.ru

Alexey E. Titov
Automatic Control Systems
South Federal University
Rostov-on-Don, Russia
alex.evgeny.titov@gmail.com

Vladimir A. Tchekhovski
Institute for Nuclear Problems,
Belarusian State University,
Minsk, Belarus,
vtchek@hep.by

Anna V. Bugakova
Student Member of IEEE
Don State Technical University,
Rostov-on-Don, Russia
annabugakova.1992@mail.ru

Abstract—We have considered results of SiPM analog interface schematic design on array chip MH2XA030. The SiPM includes current buffer, current integrator and base level recovery (BLR) circuit. We have described analog interface operation for this type of sensor in details, specified electrical circuit and computer simulation results.

Keywords—Sensors Analog Interface, Silicon Photomultipliers, SiPM, Multi-Pixel Avalanche Photodiodes, MAPD, Multi-Pixel Photon Counter, MPPC, Readout Electronic Devices, Array Chip, Integrated Circuits, Photomultipliers, Photodetectors, Circuit Simulation, Analog Circuits, Current Negative Feedback

I. INTRODUCTION

The silicon photomultipliers, called as silicon photomultiplier (SiPM), micro-pixel avalanche photodiodes (MAPD), multi-pixel photon counter (MPPC), are successfully applied in different areas of science and technology to register different types of electromagnetic radiation.

One of the most prospective approach of SiPM analog interfaces development is a development of CMOS integrated circuits' (ICs) with low power consumption level, it is achieved by switching to power supply unidirectional voltage of low value and applying signal current processing inside analog interface. The features of such IC are presented in [1-8].

The provided analog interface wide input dynamic range is essential to process signals of modern SiPMs, that include more than 1000 microcells and are consequently characterized by generated signal range that is higher than 60 dB. We have decided to design SiPM analog interfaces based on array chip MH2XA030, specially developed to provide analog radiation resistant low temperature microcircuits [9], to speed the work and reduce its costs. There were two primary activities: upgrade of two channel amplifier-discriminator Ampl-1.18 IC [10] and development of readout electronic devices with signal current processing.

To replace Ampl-1.18 IC there is a new developed amplifier ADPreamp13 [10], including transfer resistance processing amplifier, charge-sensitive processing amplifier with base level recovery (BLR) circuit and transfer

coefficient electronic adjustment. The amplifier ADPreamp13 is characterized by high level of parameters, it remains functional at gamma-radiation intensity up to 500 krad and neutron fluence impact up to 10^{13} n/cm². Unfortunately, the amplifier ADPreamp13 has a limited input dynamic range in spite of transfer coefficient electronic adjustment. It is caused by a high current-voltage transfer coefficient of input transfer resistance amplifier. The amplifier is serviceable at the minimum supply voltage of ± 3 V.

The present article purpose is to consider circuit engineering and parameters of IBUF analog interface, developed to readout signals of SiPM with wide dynamic range, the interface is implemented on array chip MH2XA030 and serviceable, when unidirectional voltage is 3.3 V.

The signal current processing inside analog interface provides the high input dynamic range. The interface is synthesized on bipolar transistors, it provides the higher level of radiation resistance, compared to implementation on CMOS transistors. The implementation on array chip provides a possibility to apply easily the developed schematic solutions in new semicustom IC.

II. DEVELOPED ANALOG INTERFACE CIRCUIT PECULIARITIES

There is an IBUF analog interface circuit, developed for MH2XA030 array chip components on Fig. 1 and Fig. 2. There is an electrical circuit of current buffer and current integrator on Fig. 1. There is a BLR circuit on Fig. 2. Note, that the like units (V_{CC} , Bias), shown on Fig. 1, 2, are connected. Their purposes are: V_{CC} – positive supply voltage; Inp – analog interface input; V_{REF} – reference voltage, that sets a d. c. voltage at input Inp ; $Iout$ – output of current, which amplitude is equal to the input one, in order to connect comparator, that registers input signal time; $Iout/10$ – output of 10 times decreased input current to connect OutBLR output of BLR circuit, if necessary; Out – current integrator output; $OutShift$ ($OutShift1$) – voltage, which sets base level at integrator output, when BLR circuit is connected (disconnected); $InpBLR$ – BLR circuit input, connected with current integrator output Out , if necessary. There are scale factors $AREA$, equal to number of parallel connected transistors on Fig. 1 and Fig.2. For example, $AREA_{13}=21$ for Q13.

The study has been carried out at the expense of the grant from the Russian Science Foundation (Project No. 16-19-00122-P).

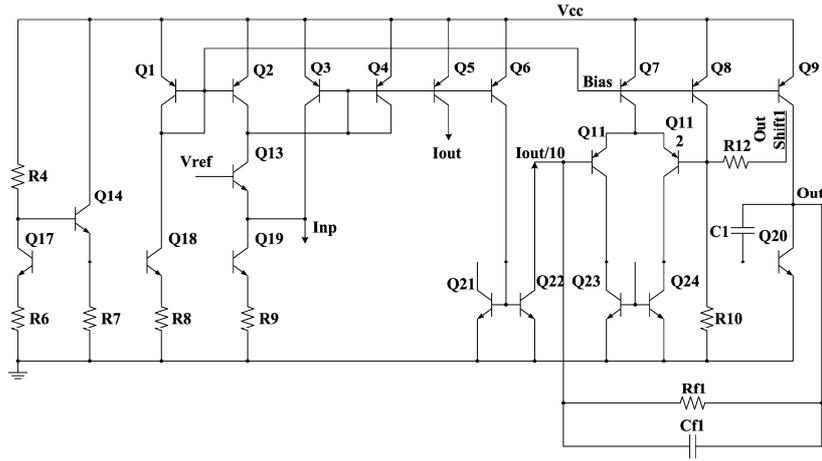


Fig. 1. Current Buffer and Current Integrator Electrical Circuit.

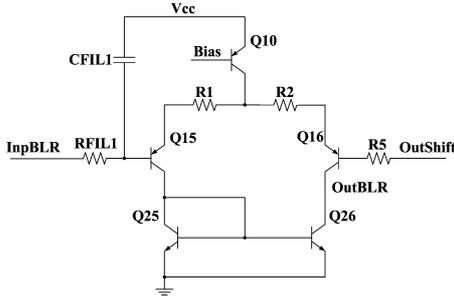


Fig. 2. Base Level Restoration Circuit.

Ibuf analog interface includes: 1) current source Q_{14} , Q_{17} , Q_{18} , R_4 , defining collector current as about $160 \mu\text{A}$, that is $I_{C18} \approx 160 \mu\text{A}$; 2) current buffer, that includes cascade with common base on low-noise transistor Q_{13} (high scale factor $AREA_{13} = 21$ provides a base area low resistance) with voltage in a form of current source Q_2 and current mirror input $Q_3 - Q_6$, one output (collector Q_3) is connected with analog interface input, the second output is connected with output I_{out} , the third one is a current integrator input via current inverter Q_{21} , Q_{22} ; 3) current integrator, that is an operational transconductance amplifier at Q_{11} , Q_{12} , Q_{20} , Q_{23} , Q_{24} , included in negative feedback circuit on R_{F1} , C_{F1} . The current source Q_8 and resistor R_{10} are applied to set a voltage on one of the operational amplifier inputs to about 300 mV . It is possible to change it, when supplying voltage from 0 to 3.3 V on output $OutShift1$; 4) BLR circuit, which can be connected parallel to integrator negative feedback circuit, if necessary, that is between $I_{out}/10$ (connected with $OutBLR$) and Out (connected with $InpBLR$). BLR circuit includes a differential amplifier Q_{15} , Q_{16} , Q_{25} , Q_{26} , the d. c. voltage is supplied to one input from $OutShift$ via resistor R_5 , a signal is supplied to the second output from low frequency filter output R_{FIL1} , C_{FIL1} , and $R_5 = R_{FIL1}$.

If we neglect transistors' base currents and consider scale factors and the specified relation of resistors' resistances $R_8 = 2R_9$, we will have the following relations of resistors' collector currents with no input signal:

$$I_{C2} = I_{C18}, I_{C19} = 2I_{C18}, \quad (1)$$

$$I_{C2} + I_{C3}AREA_4/AREA_3 = I_{C19} - I_{C3}, \quad (2)$$

$$I_{C3} = I_{C18}/(1 + AREA_4/AREA_3) = I_{C18}/1,2 = 133 \mu\text{A}, \quad (3)$$

$$I_{C5} = I_{C3}, I_{C6} = I_{C3}/AREA_3. \quad (4)$$

So, a selection of factors $AREA$ for transistors Q_1-Q_6 uniquely defines a value of their collector currents at given I_{C18} .

When collector Q_3 is connected with input Inp , there is a negative feedback circuit on transistor Q_{13} , that decreases a change of emitter current Q_{13} , when input current I_{INP} supplies, and consequently decreases an input voltage change, because almost all input current I_{INP} is provided with collector current Q_3 , $I_{INP} \approx I_{C3}$. In this case a buffer input resistance significantly decreases, and the following conditions are met:

$$I_{C5} \approx I_{INP}, I_{C6} \approx I_{INP}/10. \quad (5)$$

The current I_{INP} , decreased by 10, is converted into voltage at current integrator output Out , it allows processing large current signals of modern SiPM. When applying the developed circuit solution, it is reasonable to provide different values of input current decay in the integral circuit, for example, by forming Q_6 from several parallel transistors with emitters, connected by bus V_{CC} with keys, and different value of current integrator conversion efficiency due to switching condensers C_{F1} set, installed on the chip.

To prevent an integrator converter efficiency decrease, when a part of current impulse runs via feedback resistor R_{F1} , its resistance is selected to be the highest one. In this case the given processing tolerance of resistors' resistivity and transistor's current amplification leads to a significant spread in value of base level at output Out . It is recommended to apply BLR circuit to reduce this effect in analog interface.

Since $R_5 = R_{FIL1}$ and transistors' base currents Q_{15} , Q_{16} are practically equal, the differential amplifier of BLR compares the integrator output signal direct component (V_{OUTDC}) and voltage in $OutShift$ ($V_{OutShift}$), if they are different, it changes a loss of voltage V_{RF1} at resistor R_{F1} as long as the following condition is met: $V_{OUTDC} = V_{OutShift}$. The specified change V_{RF1} is provided due to differential amplified output current compensation of current $I_{out}/10$ part, it increases an integrator noise level almost by $\sqrt{2}$.

So, the developed BLR circuit provides a base level value control, it allows: 1) correlating the base level with AD converter output voltage, that processes IBUF signals, if necessary; 2) if Out output should be connected with comparator input, it is reasonable to connect one comparator input with bus V_{REF} and apply a base level adjustment to set comparator switch threshold.

III. GENERAL CIRCUIT SIMULATION RESULTS

It is known, that it is necessary to consider a form of SiPM output signal and parameters to provide correct readout electronic equipment simulation. We have proposed a simplified electrical equivalent circuit, given on Fig. 3, for SiPM Photonique with 516 microcells, where: N – number of firing microcells, registered photons; $Q=127.1$ fC – a charge, occurred in one microcell, when photon incomes; TD , TR , TF , PW , PER , $V1$, $V2$ – ideal current rectangular source’s parameters, that are delay time, rising and falling time, impulse duration, time, initial and final voltage correspondently.

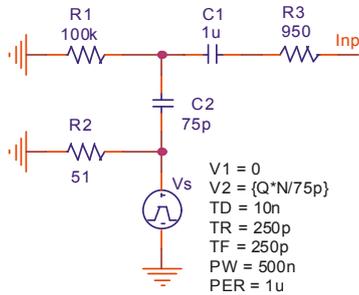


Fig. 3. SiPM Photonique Simplified Electrical Equivalent Circuit with 516 Microcells.

This equivalent circuit SiPM was applied to simulate IBUF analog interface at set number of firing microcells ($N=1, 3, 5, 10, 30, 50, 100, 300, 500$), that is at input charge $Q_{INP} = 127.1$ fC*N. The following parameters were defined, specifying equivalent capacitance SiPM C_D : 1) coefficient K_{QV} for current integrator output as relation of maximum absolute value of output voltage and maximum absolute value of input charge; 2) output voltage peak time T_P ; 3) permissible parameters adjustment ranges; 4) RMS-noise charge for current integrator output Out, reduced to analog interface input Inp. There are main simulation results on Fig. 4-8 and in Table.

The simulation has shown that:

- Input currents I_{out} , $I_{out}/10$ are significantly decreased due to negative feedback circuit (collector Q_3 connected to input Inp), applied for input transistor of current buffer Q_{13} (Fig. 4). The output current $I_{out}/10$ is not zero at $Inp = 0$, it allows decreasing current transfer coefficient $K_1 = dI_{out}/dI_{INP}$ dependence on input current. The negative feedback circuit, applied for transistor Q_{13} , increases a current transfer coefficient stability by 1.6 (Fig. 5) and provides current buffer input resistance decrease from 90.1Ω to 29Ω (Fig. 6) under otherwise equal conditions;

- Voltage level $V(Out)$ is changed from 1.28 V to 3.10 V, because a voltage of OutShift1 is changed from 0 to 3 dB without base level recovery circuit. In this case a minimum value of voltage is subject to loss in voltage at resistor R_{F1} , it can be decreased due to resistance R_{F1} decrease, if necessary;

- Base level recovery circuit allows fading at output Out, using voltage at unit OutShift. When supply voltage is 3.3 V, a recommended level shift is from 0.4 V to 2.5 V, the maximum value is defined by permissible input in-phase voltage of differential amplifier (Fig. 2), it can be increased with DA circuit update;

TABLE I. IBUF ANALOG INTERFACE PRIMARY PARAMETERS AT 3.3 V OF SUPPLY VOLTAGE

Parameter	Value
Open-circuit current consumption, mA	1.28
Input impedance, Ω	<29
Conversion coefficient K_{QV} at $C_D \approx 18$ pF, V/pC	0.03
Base level adjustment range for output Out by changing voltage in OutShift, V	from 0.4 to 2.5
Base level adjustment range for output Out by changing voltage in OutShift1, V	from 1.28 to 3.10
Peak time for output Out at $C_D \approx 18$ pF, ns	96.5
3 dB bandwidth for output Out at $C_D \approx 18$ pF, MHz	from 0.034 to 1.84
RMS-noise charge for output Out, reduced to analog interface input Inp, at $C_D \approx 18$ pF with connected (disconnected) BLR circuit, fC	58.23 (35.94)

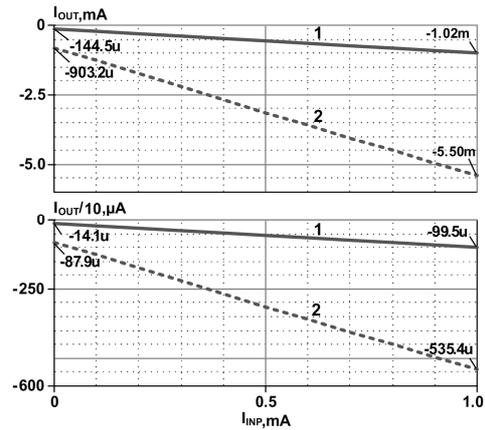


Fig. 4. Input Currents I_{out} , $I_{out}/10$ Dependence on Input Current I_{INP} : 1 – with Connected Collector Q_3 with Input Inp; 2 – with Disconnected Collector Q_3 with Input Inp.

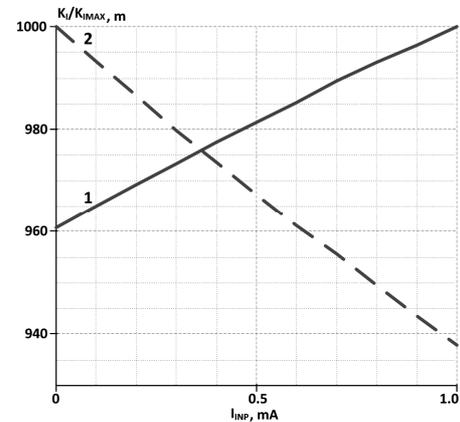


Fig. 5. Output Current Transfer Standard Rate $I_{out}/10$ Dependence on Input Current: 1 – with Connected Collector Q_3 with Input Inp; 2 – with Disconnected Collector Q_3 with Input Inp.

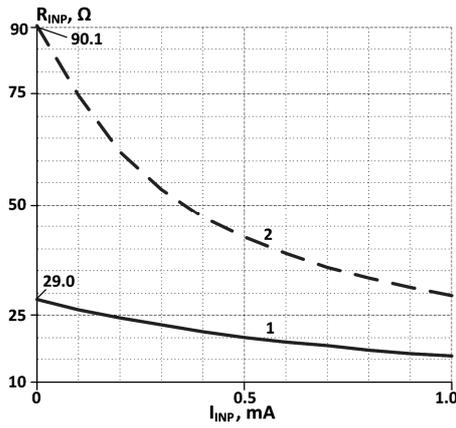


Fig. 6. Buffer Input Resistance R_{INP} Dependence on Input Current I_{INP} : 1 –with Connected Collector Q_3 with Input Inp; 2 – with Disconnected Collector Q_3 with Input Inp.

• If we compare Fig. 7 and Fig. 8, we may say, that BLR circuit provides a negligible change of base level at $\pm 20\%$ range of integral resistors' resistance (Fig. 8), but K_{QV} is changed. It occurs due to resistor R_{F1} resistance influence on impulse amplitude at output Out. Moreover, BLR circuit compensates input current up to $200\ \mu A$ influence on base level;

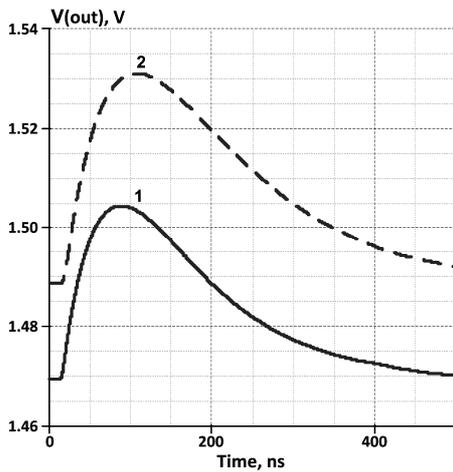


Fig. 7. Voltage Impulse at Generator Output Out without BLR Circuit for 10 Firing Microcells SiPM ($Q_{INP} = 1.271\ pC$) at resistors' resistance range: 1 – $0.8R$; 2 – $1.2R$.

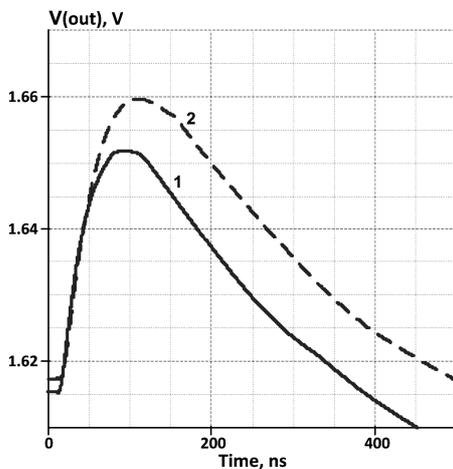


Fig. 8. Voltage Impulse at Generator Output Out with Base Level Recovery Circuit for $Q_{INP} = 1.271\ pC$ at Resistors' Resistance Range: 1 – $0.8R$; 2 – $1.2R$.

• Developed analog interface ensures safe operation at gamma-radiation intensity up to 1 Mrad, influence of integral neutron flux up to $10^{14}\ n/cm^2$, provides stable registration of photon number from 1 to 500 by SiPM Photonique, as it generates a minimum charge equal to 127.1 fC, when one photon incomes, it is 3.54 times higher than analog interface noise level without BLR circuit.

I. CONCLUSION

We have developed IBUF analog interface based on array chip MH2XA030 for SiPM with wide dynamic range, including current buffer, current integrator and BLR circuit.

When unidirectional supply voltage is 3.3 V, the analog interface is characterized by the following main parameters: input resistance is $29\ \Omega$, current consumption is 1.28 mA, output impulse peak time is about 100 ns.

BLR circuit provides a negligible change of base level at $\pm 20\%$ range of integral resistors' resistance, compensates input current up to $200\ \mu A$ influence on base level, but simultaneously increases noise level by 1.6.

The developed analog interface ensures safe operation at gamma-radiation intensity up to 1 Mrad, influence of integral neutron flux up to $10^{14}\ n/cm^2$ and may be applied in SiPM signal processing multichannel microcircuits.

REFERENCES

- [1] F. Corsi, M. Foresta, C. Marzocca, G. Matarrese, A. Del Guerra, "ASIC development for SiPM readout," 2010 IEEE Nuclear Science Symposium Conference Record (NSS/MIC), Sept. 23–26, 2008. DOI: 10.1109/NSSMIC.2010.5874056
- [2] Y. Bocharov and V. Butuzov, "An analog front-end ASIC with programmable gain and timing for silicon photomultiplier arrays," 2016 IEEE SIBCON, Moscow, 2016, pp. 1-4. DOI: 10.1109/SIBCON.2016.7491685
- [3] D. Meier, et al., "SIPHRA 16-Channel Silicon Photomultiplier Readout ASIC," IDEAS to Proc. AMICSA&DSP 2016, 12-16 June 2016, Sweden, pp. 1-7. DOI: 10.13140/RG.2.1.1460.8882.
- [4] S. Dey, et al., "A CMOS ASIC design for SiPM arrays 2011 IEEE Nuclear Science Symposium Conference Record, Valencia, 2011, pp. 732-737. DOI: 10.1109/NSSMIC.2011.6154092
- [5] P. Dorosz, M. Baszczyk, W. Kucewicz and Ł. Mik, "Low-Power Front-End ASIC for Silicon Photomultiplier," in IEEE Transactions on Nuclear Science, vol. 65, no. 4, pp. 1070-1078, April 2018. DOI: 10.1109/TNS.2018.2816239
- [6] F. Corsi, M. Foresta, C. Marzocca, G. Matarrese and A. Del Guerra, "A self-triggered CMOS front-end for Silicon Photo-Multiplier detectors," 2009 3rd International Workshop on Advances in sensors and Interfaces, Trani, 2009, pp. 79-84. DOI: 10.1109/IWASI.2009.5184772
- [7] H. Chen et al., "A dedicated readout ASIC for Time-of-Flight Positron Emission Tomography using Silicon Photomultiplier (SiPM)," 2014 IEEE NSS/MIC, Seattle, WA, 2014, pp. 1-5. DOI: 10.1109/NSSMIC.2014.7431045
- [8] Y. Chen, Z. Deng and Y. Liu, "Development of a Multi-Channel ASIC with Individual Energy and Timing Digitization for SiPM Detectors for TOF-PET Applications," 2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Atlanta, GA, 2017, pp. 1-4. DOI: 10.1109/NSSMIC.2017.8532718
- [9] O. V. Dvornikov, et al., "Basic Parameters and Characteristics of the Op-Amp Based on the BiFet Array Chip MH2XA030 Intended for the Design of Radiation-Hardened and Cryogenic Analog ICs," 2018 XIV IEEE APEIE, Novosibirsk, 2018, pp. 200-207. DOI: 10.1109/APEIE.2018.8545562
- [10] O. V. Dvornikov, et al., "Integrated microcircuit for registration of silicon photomultiplier signals," J. Instruments and equipment of the experiment, 2014, No. 1, pp. 66-71.

A novel technique for atomic instructions functional verification using lock contention analysis

Chibisov Peter
computing systems development
dept.
SRISA RAS
Moscow, Russia
chibisov@cs.niisi.ras.ru

Grevtsev Nikita
computing systems development
dept.
SRISA RAS
Moscow, Russia
ngrevtsev@cs.niisi.ras.ru

Abstract—Nowadays, it is widely acknowledged that symmetric multi-processing (SMP) must use a set of synchronization mechanisms to achieve the results, which are free of race conditions and therefore predictable. Using atomic operations to access shared memory regions in SMP systems has been proven to be the basic method to prevent data corruption while implementing in software such primitives as mutual exclusion, spinlock, thread execution barrier. Contemporary architectures provide different kinds of atomic operations, for example, LL and SC instructions on MIPS are used together to guarantee atomic read-write accesses to shared memory.

However, synchronization mechanisms such as critical sections used to assure exclusive access to critical resources and data structures are well-known potential performance bottlenecks in multithreaded applications. There are numerous approaches proposing a variety of methods to analyze and reduce these problems in software or in hardware. These approaches quantify the execution overhead of synchronization mechanisms or assess the impact these primitives have on the completion time of multithreaded applications.

A key finding in this paper is that all these researches and experience mentioned can be used to increase the coverage of atomic operations functional verification. Moreover, our experience has shown that atomic operations functional verification is a rather time-consuming and labor-intensive process because atomic operations can not be verified with the help of stochastic test generation methods due to their unpredictable nature. This is evidenced by the lack of sufficient information on the topic, which can mean that atomic operations could have been successfully verified by traditional methods. However, one can find at least two CPU's errata references where some issues concerning atomic operations misbehaviors have been listed.

To the best of our knowledge, our synchronization mechanisms analysis is the first method that focuses not on performance issues, but on functional verification of atomic operations instructions, which provide a basis for these mechanisms.

Keywords—lock contention, cache contention, atomic instructions, functional verification, random test generation, PARSEC benchmark, lock torture, test coverage, simulator

The study was provided by the support of the state fundamental research program No. 0065-2019-0004.

I. INTRODUCTION

Our experience has shown that atomic operations functional verification is a rather time-consuming and labour-intensive process because atomic operations can not be verified with the help of stochastic test generation methods due to their unpredictable nature. Instruction Set Simulator (ISS), which is generally used to obtain expected responses for the test being executed, is not applicable. In the paper, we adopted lock contention analysis and prediction approaches to create a novel technique for atomic instructions analysis the main purpose of which is to increase the coverage of atomic operations functional verification.

It is commonly recommended to avoid data conflicts while writing concurrent programs [1]. For example, false sharing is a well-known problem for multithreaded applications that can radically degrade both performance and scalability [2], [3], [4]. Undoubtedly, some form of synchronization between the threads, such as a barrier or a critical section, is needed and, therefore, data conflicts while acquiring lock cannot be avoided (this is called true sharing). Using atomic operations to access shared memory regions in SMP systems are the basic method to prevent data corruption. Atomic operations are the lowest level synchronization primitives. They are used as building blocks for higher level constructs, like, for example, mutual exclusion, spinlock, and thread execution barrier (Fig. 1). Synchronization mechanisms such as critical sections suffer from contention [5].

There are numerous approaches proposing a variety of methods to analyze and reduce these problems in software or in hardware. These approaches quantify the execution overhead of synchronization mechanisms or assess the impact these primitives have on the completion time of multithreaded applications [6] - [15]. However, it is reasonable to create test cases with forced shared data areas between threads intentionally if we want to test the interaction between cores. A key finding in this paper is that all previous researches and experience can be used to increase the coverage of atomic operations functional verification.

The rest of the paper is organized as follows. Section 2 describes basic concepts of atomic operations and their function at low-level and from a programmer's point of view. Moreover, we describe why existing methods are not suitable for the verification of atomic operations and solution to increase the coverage of atomic operations is proposed.

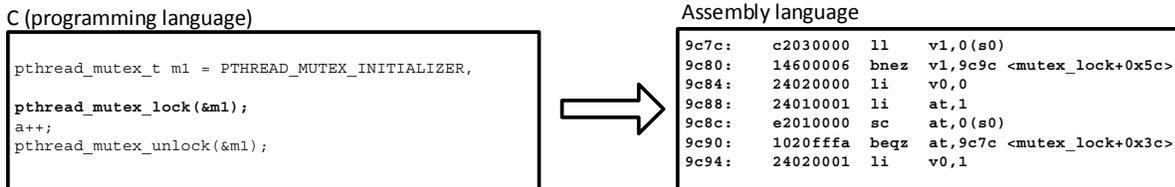


Fig. 1. Example of atomic operations in lock code

Section 3 presents the experimental evaluation and discusses how our knowledge about lock contention can help us understand how to extend random tests generation techniques for atomic instructions verification. Section 4 discusses our novel technique for increasing atomic operations contention in random test cases. Section 5 concludes the paper.

II. BASIC CONCEPTS

According to MIPS Architecture documentation [16], the LL (Load Linked) and SC (Store Conditional) instructions provide primitives to implement atomic read-modify-write (RMW) operations for memory locations. When an LL is executed it starts an active RMW sequence replacing any other sequence that was active. The RMW sequence is completed by a subsequent SC instruction that either completes the RMW sequence atomically and succeeds, or fails. Executing LL on one processor core does not cause an action that, by itself, causes an SC for the same block to fail on another processor core. The SC fails if any of the following events occurs between the execution of LL and SC: a coherent store is completed by another processor core or coherent I/O module into the block of physical memory, containing the word or “exception return” instruction is executed.

LL and SC are used to atomically update memory locations, as shown below:

```

L1: LL T1, (T0)      # Load Linked - load counter
    ADDI T2, T1, 1   # increment
    SC T2, (T0)      # Store Conditional - try to store,
                    # checking for atomicity
    BEQ T2, 0, L1    # if not atomic (0), try again
    NOP              # branch-delay slot

```

The program code shown above is an example of a basic part, which is used to construct such primitives as mutual exclusion, spinlock and thread execution barrier.

The testing process begins with the creation of simple hand-written tests in assembly language aimed at checking essential behavior of the atomic instructions. Clearly, even the simplest example of a hand-written test should take into account all relevant factors, which can result in errors and inconsistencies. Of course, it is impossible to conduct an exhaustive search through all possible test combinations in manual way. Therefore, the first question is how to conduct a heuristic search and create automated or partially automated test generation system. The second question is even more fundamental: atomic operations can not be verified with the help of stochastic test generation methods due to their unpredictable nature. That means that Instruction Set Simulator (ISS), which is generally

used to obtain expected responses for the test being executed, is not applicable. Therefore, we have had to use one of the self-checking approaches.

We use simple hand-written assembly language tests with the internal self-check at the early stages of the RTL-model design cycle. The simplest examples of parallel program algorithms with data races, which can be implemented without operation system libraries support, are arrays addition and linked lists operations [1]. These algorithms were chosen for the following reasons: 1) they represent fragments of real computing tasks; 2) they can be manually divided into the number of parallel threads; 3) one can easily designate a critical sections to protect data variables that are linked and cannot be split between threads; 4) addresses of locks (i.e. of atomic instructions considered) may be intentionally mixed with data structures variables to increase contention. Furthermore, we have found that test generation of such tests should be partially or fully automated for successful functional verification of LL and SC instruction pair. True sharing and false sharing situations may also be mentioned among the relevant factors that have strong impact on the execution of LL and SC instructions.

In this paper, we propose a solution to increase the coverage of atomic operations within the period of RTL-model functional verification. The solution is largely based on results of previous studies [7], [9] dealing with issues of hardware contention (for example, cache and memory), software contention (for locks and thread barriers), parallelization overhead, and work load imbalance [17]. Mainly, we focus on understanding of lock contention. In a parallel program, the use of shared data is typically protected by locks to guarantee exclusive access. If several threads try to acquire the same lock, only one thread at a time can succeed and the others must typically wait instead of doing useful work [7]. This is commonly referred to as contention of the lock [7], [8].

Moreover, synchronization mechanisms such as critical sections used to assure exclusive access to critical resources and data structures are well-known potential performance bottlenecks in multithreaded applications [10]. We analyzed numerous approaches proposing a variety of methods to understand and reduce these problems in software or in hardware. These approaches quantify the execution overhead of synchronization mechanisms or assess the impact these primitives have on the completion time of multithreaded applications. In contrast, instead of improvement processor core performance, we use the concept of lock contention (this is also true for “thread contention” or even “cache contention” terms) to create new technique for increasing quality of functional verification of atomic operations instructions (such as Load

Linked and Store Conditional instructions from MIPS processor's ISA).

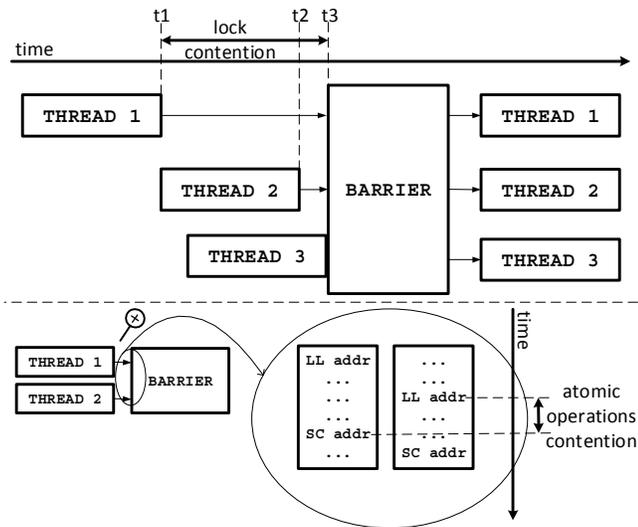


Fig. 2. Lock contention vs atomic contention on the example of barrier

It should be emphasized that the lock contention concept is directly related to atomic instructions (Fig. 2). However, from a programmer's point of view, lock contention is a condition that a) limits the amount of parallelism by serializing accesses to protected shared data or b) associated with the idleness in the thread execution. Normally, these conditions are referred to as a performance degradation issues, and it is considered important to minimize the number of atomic operations, which are required during critical sections [9]. In spite of the fact that it is recommended to avoid data conflicts while writing concurrent programs and kernel drivers, we consider it essential, as a part of this work, to create random test cases with forced shared data areas between threads intentionally to test the interaction between atomic instructions in a shorter period of time.

From a more general point of view, our contention definition differs from universally accepted. We need to create atomic operations contention in order to verify their functionality

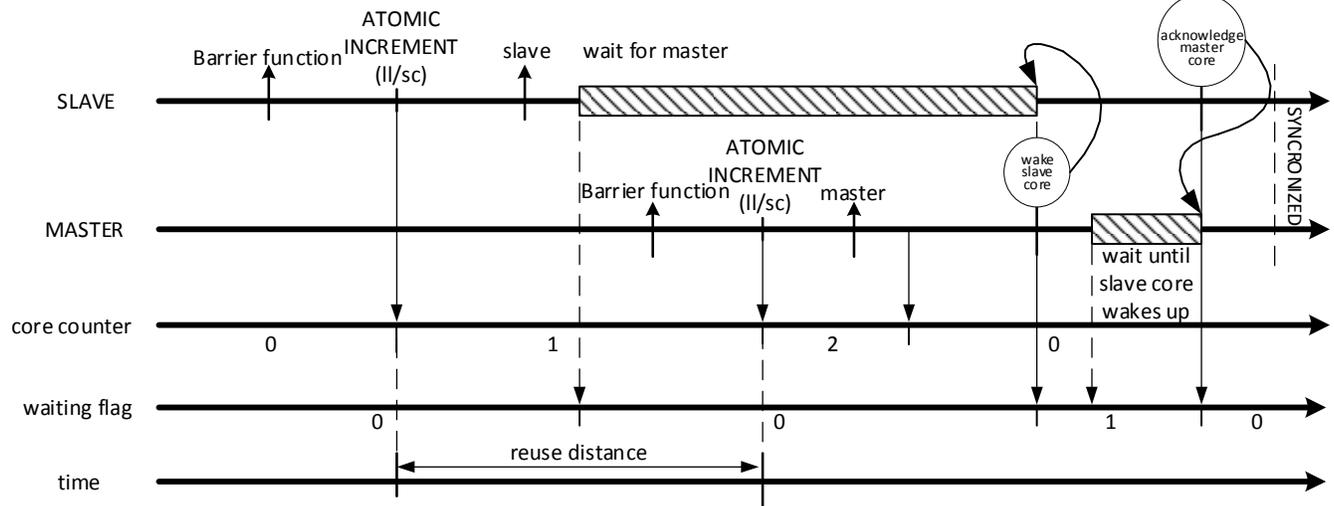


Fig. 3. General view of barrier function

exhaustively. Let us define this type of contention (for simplicity let us take two cores only) as a coincidence of two simultaneously executed pairs of Load Linked and Store Conditional instructions on different processor cores if they use the same physical address and one of the Store Conditional instructions fails.

Our experience shows that the amount of LL-SC contention tends to be by several orders of magnitude less likely to occur than overall cache or lock contention. This can be explained by the fact that LL-SC contention arises as a consequence of cache contention (Fig. 3), but SC has much less probability to fail. In this respect, it is also worth noting that the number of unique memory references made between two atomic accesses to the same memory location is of particular interest [18]. This method called the reuse distance analysis can be effective to measure data locality of atomic operations [19].

III. EXPERIMENTAL EVALUATION

This section answers the following questions:

- 1) What is the relationship between the lock contention and atomic operations contention in terms of execution time?
- 2) What is the probability of atomic operations execution while running some representative benchmarks?
- 3) How to evaluate Store Conditional failure rate?
- 4) How to use the information obtained in order to improve functional verification of Load Linked and Store Conditional instructions?

To answer these questions, real applications from state-of-the-art multithreaded program benchmark suite PARSEC 2.1 are used [20]. We run a representative part of each benchmark on MIPS Instruction Set Simulator to get sufficient instruction trace. It has been taken into account that for each benchmark there is a Region of Interest (ROI) [21], which indicates what part of the benchmark executes in parallel. The ROI is also important for ensuring that results obtained using simulation inputs represent real program behavior.

TABLE I. THE EVALUATION RESULTS FOR LL AND SC INSTRUCTIONS BEHAVIOR

Test name		Total Reads (M)	Total Writes (M)	Total LL executed	Total SC executed	Unique PCs of LLs	LL unique access addresses	Private/Shared LL addresses	SC Failed (LL-SC Broken)
PARSEC	streamcluster	55.08	0.65	6,695	6,033	18	17	13/4	122
	fluidanimate	54.83	10.46	219,779	219,770	37	3449	1,914/1,535	4
	bodytrack	38.49	12.28	129,893	129,866	123	187	141/46	0
	blackscholes	3.51	1.46	91,452	91,450	53	82	82/0	0
	ferret	47.80	15.96	23,677	23,515	23,666	637	549/66	5
LTP: lock torture	spin_lock	40.17	17.20	1,563,381	1,559,146	54	82	44/38	591
	spin_lock_irq	39.12	17.05	1,493,647	1,489,073	19	42	23/19	483
	rw_lock	17.99	3.10	478,720	476,367	30	40	23/17	2
	rw_lock_irq	15.31	4.76	865,469	864,679	28	40	32/8	0
	mutex_lock	8.78	2.67	530,696	530,688	30	54	42/12	0
	rwsem_lock	7.60	2.31	459,437	459,437	18	36	32/4	0
	rtmutex_lock	34.29	17.05	2,028,869	1,886,839	13	21	18/3	0
	ww_mutex_lock	14.75	4.47	892,719	892,717	24	35	34/1	0
	percpu_rwlock	12.12	3.68	732,757	732,756	23	45	26/19	0

By simulating only the ROI with simmedium input set (which is suitable for microarchitectural studies with simulators [22]), we have been able to reduce simulation time.

Moreover, we evaluate our strategy of measuring SC fail rate on Linux Kernel Lock Torture Test [23], which can be run as a part of Linux Test Project (LTP) [24]. This torture test consists of a number of kernel threads, which acquire the lock and hold it for specific amount of time, thus simulating different critical region behaviors. The amount of contention on the lock can be simulated either by enlargement of this critical region hold time and/or by creating more kthreads [23].

Firstly, we have consistently studied summary of the key characteristics of PARSEC benchmarks, especially breakdown of synchronization primitives [20], [25], [26], [27]. Secondly, we have selected 5 benchmarks and some other user applications reported to have the most contended locks. Thirdly, we have done simulations and obtained the required data. This allows us to get a deep insight into the structure of synchronization primitives. In addition, this knowledge about lock contention can help us understand how to extend random tests generation techniques for atomic instructions verification.

We run experiments on a 2 core MIPS simulator running mips64 Debian GNU/Linux 9 “stretch”. We use gcc-6.3.0 with “-O3” optimization with Linux kernel 4.15. All experiments with PARSEC benchmark use simmedium input sets. We simulate only the ROI of each benchmark and use instruction trace, which contains all the information needed from 10⁸ processor instructions on each core.

We have obtained load and store instructions ratio, atomic instructions (LL and SC) ratio, unique atomic instructions executed on one core only ratio (their Program Counters, PCs), shared and private addresses ratio for atomic instructions, and

we have examined Store Conditional failed events. Shared memory locations are measured with cache line granularity.

The evaluation results for LL and SC instructions behavior are shown in Table 1.

It is important to explore all typical blocks of program code containing atomic instructions and types of operations that caused SC failure. However, the probability of SC instructions failure observed in complex real-world applications being run on the simulator shows that race conditions and SC failures are extremely rare for such a shared addresses correlation. To explain the reason why Store Conditional failed events are so rare at runtime, we illustrate a representative execution of the typical benchmark part in Fig. 4.

As we can see, even though there are many examples of shared memory locations being jointly used, they are used during different time periods. As a result, LL-SC contention is low. However, quite often, it happens that one thread (process) works with some data on core 1 and, after hardware interrupt, the operating system chooses to change the affinity of this process and assigns it to core 2. This context switching will cause many atomic instructions to be executed on both cores, however, LL-SC contention may still be low.

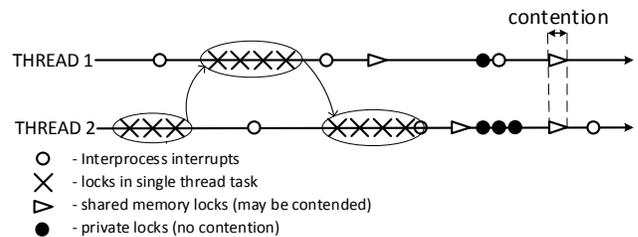


Fig. 4. A representative execution diagram of the typical benchmark

Our strategy, on the contrary, is designed to increase LL-SC contention intentionally while generating and running random verification tests for RTL-model of the multicore processor.

How this is done is explained in the next section.

IV. A TECHNIQUE FOR INCREASING ATOMIC OPERATIONS CONTENTION IN RANDOM TEST CASES

After analyzing the most contended tests behavior, we came to the conclusion that in order to effectively verify this area of the microprocessor design, we need to solve 3 main tasks:

- Reuse distance significant reduction (temporal locality) [18].
- Percentage of shared addresses increase (spatial locality).
- Diversification of LL-SC cycles.

All these tasks can be solved with the help of directed self-checking random test generation.

It is important to realize that errors in the RTL-model behavior are usually associated with a combination of many factors, such as all levels caches hits or misses, different types of registers dependencies, the behavior of the branch prediction unit, etc. From this perspective, different types of atomic operation's cycle structure result in various types of pipeline unique dependencies during execution, which can result in functional errors. Therefore, one of the test coverage metrics is related to LL-SC cycles variability. Nevertheless, it should be noted that fully randomly (an arbitrarily) generated LL-SC loop may be either inherently improper (it may cause a live-lock), or the result of its execution is unpredictable.

The obtained unique combinations of LL-SC cycles (see Table 1) can be used for further testing process. We assume that set of tests discussed above covers the majority of ways of using the atomic operations. All the combinations obtained can be classified according to their structure (in terms of the instructions types and their dependencies).

To cover all the remaining possible situations, the resulting set of test situations can be formalized and presented in the form of pseudocode shown in Fig. 5. The cycle always consists of instructions LL and SC accessing to the same address and the branch instruction depending on the SC instruction execution result. The loop can also contain an internal LL-dependent branch instruction that may restart or interrupt the loop execution. Space between LL and SC can be filled with an arbitrary number of random instructions with some limitations.

In this case, some restrictions on the cycle generation must be set:

- There can be no other memory access operations inside the loop body according to the MIPS Architecture documentation;
- Atomic operation cycle in the test case should be constructed in a way to achieve a predictable final state of the registers and memory state regardless of how many times the loop is repeated (SC failure must lead to reexecution of all LL-SC cycle);

- Guaranteeing the feasibility of the LL-SC cycle regardless of the memory state and another core behavior;
- LL-dependent branch should not cause infinite loop.

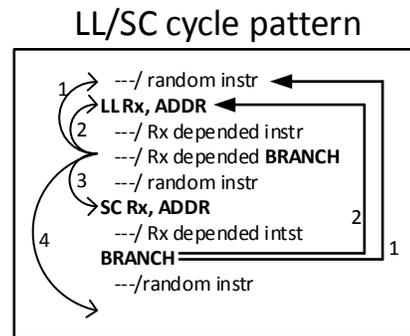


Fig. 5. Formalized LL-SC cycle

To direct the generator toward the interesting events (particular areas in the design or some particular scenario of atomic instructions execution), testing knowledge is embedded into the generator [28]. Such situations are obviously absurd and meaningless from programmer's point of view because they cannot occur in the real applications. At the same time, they have increased complexity and diversification, which aimed to improve the quality of generated test cases. The combination of all LL-SC cycles generated by the proposed models provides 100% coverage of LL-SC types.

In order to ensure atomic operations correct function, it is necessary to check whether SC failure is correct depending on various types of coherent stores at any time moment. Since it is impossible to solve this issue with the help of combinatorial tests, a pseudorandom directed test generator with a sufficiently large randomization can serve as a cause of random LL-SC atomicity failures. We have extended our current test generator Ristretto [29] to create test cases with random combinations of load/store and atomic instructions to generate necessary stimulus. Moreover, threads created by the generator can share some memory resources to initiate interactions and coherency transactions on the system bus. Furthermore, we can direct the testing process by varying the false sharing probability, the memory regions minimum and maximum sizes, the frequency of macros occurrence in the test. The pseudo-random LL-SC pairs, obtained in the previous step, are added to the test template in the form of macros and have the same memory distribution as usual load/store instructions.

Since each thread has access to its own bytes in a cache line (and in a memory area) during a subtest, the final state of the memory segment LL-SC loop operates with stay deterministic [30]. At the same time, memory accesses from different threads can belong to matching cache lines and test sequences of memory accesses have random nature. This leads to the fact that the probability of SC failure depends on temporal locality and spatial locality, which are set in the template. Therefore, the total LL-SC contention rate depends on pseudo random generator quality and testing knowledge of memory alignment.

An example of the test obtained by the presented method, as well as the memory distribution between the threads, is shown in Fig. 6. Evidently, every atomic macro in such a case has very low reuse distance due to large amount of memory requests to the same cache-line from the other core.

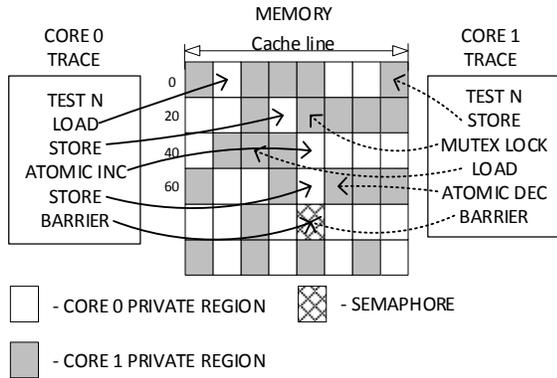


Fig. 6. Memory distribution between the threads

One of the important consequences of the injection of atomic macros is the quality improvement of the generator itself. A serious problem of random load/store streams is the problem of memory degradation, which leads to bug masking [31]. Fortunately, one can use “smart” atomic macros, which attempt to recover memory only after encountered a zero value. We guarantee that any LL-SC cycle type from the most contended applications is covered by generated test cases in different execution contexts.

By evaluating generated random test’s characteristics in the same manner as it has been done for Linux applications in Section 3 we can compare temporal and spatial locality with the data obtained in Table I. The results for LL and SC instructions behavior for generated tests are shown in Table 2.

TABLE II. RESULTS FOR GENERATED TESTS

Test name	Total Writes (M)	Total LL/SC executed (M)	Private/Shared LL addresses	SC Failed (M)
Ristretto	12.48	4.58 / 2.77	6 / 86	0.15
Array ADD	9.96	2.43 / 2.43	0 / 85907	0.008

According to these data, we succeeded in increasing ratio of LL and SC instructions, the probability of SC failure, diversification of LL-SC cycles; as well as we decrease reuse distance between shared addresses to generate high quality stress tests. These results indicate that we have been able to increase LL-SC contention while generating and running random verification tests for RTL-model of the multicore processor. One of the consequences of this remarkable improvement in the random test generation process is so-called bug rate, and, in general, the ability to discover hard-to-find bugs in a limited amount of time.

All errors found by this method can be classified into several groups:

- 1) Critical errors:
 - SC instruction reported success but did not write the value in memory.
 - SC instruction reported failure but wrote the value in memory.
 - SC instruction did not recognize a coherent store from another core and write the value in memory.
- 2) Non critical errors:
 - SC instruction reported failure without a reason (it did not cause an error but degraded performance).

Note that these types of errors are not typically directly related to the characteristics measured and shown in Table 1 or in Table 2.

The LL-SC cycle formalized and described above implements heuristics to solve the combinatorial explosion problem. However, we expect that any types of atomic operations errors can be found by proposed technique due to random nature of generated test cases provided sufficient number of tests are run.

V. CONCLUDING REMARKS

We focus on the concept of lock contention to increase temporal locality and spatial locality of atomic operations in order to increase the probability of lock contention, which is directly connected to atomic instructions. Our strategy, contrary to many previous researches, is based on LL-SC contention which is intentionally increased while generating and running random verification tests for RTL-model of the multicore processor. We prove that the proposed novel technique will cover all highly sophisticated scenarios of atomic instructions usage, which can be found in real-world applications.

The discussed approach was successfully applied to the verification of the RTL-model of dual-core microprocessor with SMP developed in SRISA RAS. The method made it possible to find the majority of memory consistency bugs, pipeline stalls, livelocks, deadlocks and other atomic operations misbehaviors. The advantage of our technique is that there is no need to change verification strategy and process to adjust to a new project design. Additionally, we can find bugs automatically because of stochastic nature of the generator.

We plan to extend our work by more detailed analysis of atomic operation contention in different kinds of for parallel workloads by using shared footprint metric for measuring locality and memory behavior of programs. We are going to use two new metrics called shared footprint and sharing ratio [32] to capture the amount of active datasharing in a threaded execution for this reason. In future work, we will also focus on the coverage of the test cases with atomic instructions generated using the proposed technique.

ACKNOWLEDGMENT

We would like to thank Aleksey Kuleshov for help with Instruction Set Simulator setup, Dmitriy Trubitsyn for scripting and Pavel Zubkovsky for his valuable suggestions and feedbacks that helped improve this paper.

REFERENCES

- [1] M. Herlihy and N. Shavit. *The Art of Multiprocessor Programming*, Revised Reprint. Elsevier, 2012.
- [2] T. Liu et al., "PREDATOR: Predictive false sharing detection", PPOPP 2014.
- [3] Tongping Liu, Xu Liu, Cheetah: detecting false sharing efficiently and effectively, Proceedings of the 2016 International Symposium on Code Generation and Optimization, March 12-18, 2016.
- [4] Zhao, Q., Koh, D., Raza, S., Bruening, D., Wong, W.-F., Amarasinghe, S. Dynamic Cache Contention Detection in Multi-threaded Applications. In Proceedings of the international conference on Virtual Execution Environments, VEE'11, pages 27–38, 2011.
- [5] A. Starke. *Locking in os kernels for smp systems*. Citeseer, 2006.
- [6] Xu, C., Chen, X., Dick, R.P., Mao, Z.M.: Cache Contention and Application Performance Prediction for Multi-Core Systems. In: Proceedings of the 2010 IEEE International Symposium on Performance Analysis of Systems and Softwares IEEE ISPASS 2010, White Plains NY, USA, pp. 76–86.
- [7] X Pan, J Lindén, B Jonsson Predicting the Cost of Lock Contention in Parallel Applications on Multicores using Analytic Modeling - MCC12, 2012.
- [8] R. Gu, G. Gin, L. Song, L. Zhu, and S. Lu, "What change history tells us about thread synchronization", in FSE, August 2015.
- [9] N. R. Tallent, J. M. Mellor-Crummey, and A. Porterfield, "Analyzing lock contention in multithreaded applications," in PPOPP, February 2010.
- [10] G. Chen, P. Stenstrom Critical lock analysis: Diagnosing critical section bottlenecks in multithreaded applications. In Proceedings of Supercomputing: the International Conference on High Performance Computing, Networking, Storage and Analysis (SC), pages 71:171:11, Nov. 2012.
- [11] H. Ding, X. Liao ; H. Jin ; X. Lv ; R. Guo Reducing lock contention on multi-core platforms. In Proceedings of 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), 2014.
- [12] Gramoli, V. More than you ever wanted to know about synchronization: synchrobench, measuring the impact of the synchronization on concurrent algorithms. In Proceedings of 20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP 2015, pp. 1–10.
- [13] Kempf S., Veldema R., Philippsen M. (2013) Compiler-Guided Identification of Critical Sections in Parallel Code. In: Jhala R., De Bosschere K. (eds) *Compiler Construction*. CC 2013. Lecture Notes in Computer Science, vol 7791. Springer, Berlin, Heidelberg.
- [14] S. Dutta, S. Manakkadu, and D. Kagaris, "Classifying performance bottlenecks in multi-threaded applications," in MCSoc 2014, September 2014.
- [15] Sahelices B., Ibáñez P., Viñals V., Llabería J.M. (2009) A Methodology to Characterize Critical Section Bottlenecks in DSM Multiprocessors. In: Sips H., Epema D., Lin HX. (eds) *Euro-Par 2009 Parallel Processing*. Euro-Par 2009. Lecture Notes in Computer Science, vol 5704. Springer, Berlin, Heidelberg.
- [16] MIPS64™ Architecture For Programmers, V.2: The MIPS64™ Instruction Set Reference Manual, Revision 6.04, MIPS Technologies Inc., 2015, 551 pp.
- [17] Jungju Oh, Christopher J. Hughes, Guru Venkataramani, and MilosPrvulovic. 2011. LIME: A Framework for Debugging Load Imbalance in Multi-threaded Execution. In Proceedings of the 33rd International Conference on Software Engineering (ICSE). 201–210.
- [18] K. Beyls, E. D'Hollander Reuse Distance as a Metric for Cache Behavior. In Proceedings of the IASTED International Conference on Parallel and Distributed Computing and Systems, IASTED, 2001. p.617-622.
- [19] R. Hemani, S. Banerjee, A. Guha, "Accord: An analytical cache contention model using reuse distances for modern multiprocessors", 21st Annual International Conference on High Performance Computing Student Research Symposium (HiPC SRS), December 2014.
- [20] Bienia, S. Kumar, J. P. Singh, K. Li, The parsec benchmark suite: Characterization and architectural implications, in: Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques, 2008.
- [21] G. Southern and J. Renau. Analysis of PARSEC workload scalability. In 2016 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), pages 133–142, April 2016.
- [22] A Memo on Exploration of SPLASH-2 Input Sets PARSEC Group Princeton University, 2011.
- [23] Kernel Lock Torture Test Operation [Electronic resource] <https://www.kernel.org/doc/Documentation/locking/locktorture.txt>
- [24] Modak, Singh, YAMATO Putting LTP to Test - validating the Linux Kernel and test cases. Proceedings of the 2009 Montreal Linux Symposium, July 2009, Montreal, Canada.
- [25] Baier C. et al. (2012) Waiting for Locks: How Long Does It Usually Take?. In: Stoelinga M., Pinger R. (eds) *Formal Methods for Industrial Critical Systems*. FMICS 2012. Lecture Notes in Computer Science, vol 7437. Springer, Berlin, Heidelberg.
- [26] Major Bhadauria, Vincent M. Weaver, Sally A. McKee: Understanding PARSEC performance on contemporary CMPs. In IEEE International Symposium on Workload Characterization, IISWC 2009: 98-107.
- [27] PARSEC benchmarks: Benchmarking Modern Multiprocessors. Christian Bienia. Ph.D. Thesis. Princeton University, January 2011.
- [28] IBM: Katz, Yoav, Rimon, Michal, Ziv, Avi and Shaked, Gai. "Learning microarchitectural behaviors to improve stimuli generation quality." Paper presented at the meeting of the DAC, 2011.
- [29] Ristretto: random test generator for multicore microprocessor cache coherence verification A.V. Smirnov, P.A. Chibisov SELECTED ARTICLES of the VIII All-Russia Science&Technology Conference MES-2018 Part 1 "in press"
- [30] Multicore Processor Models Verification in the Early Stages N.A. Grevtsev, P.A. Chibisov SELECTED ARTICLES of the VIII All-Russia Science&Technology Conference MES-2018 Part 1 "in press"
- [31] Doowon Lee, Tom Kolan, Arkadiy Morgenshtein, Vitali Sokhin, Ronny Morad, Avi Ziv, Valeria Bertacco: Probabilistic bug-masking analysis for post-silicon tests in microprocessor verification. DAC 2016: 24:1-24:6.
- [32] H Luo, X Xiang, C Ding: Characterizing active data sharing in threaded applications using shared footprint. In Proceedings of the The 11th International Workshop on Dynamic Analysis, WODA '2013.

Calculation of Phase Center of Arbitrary Electromagnetic Radiation Sources in Near Field Zone

Nikolay Anyutin
NIO-1
VNIIFTRI
Mendeleevo, Russia
anyutin@vniiftri.ru

Ivan Malay
NIO-1
VNIIFTRI
Mendeleevo, Russia
malay@vniiftri.ru

Alexey Malyshev
NIO-1
VNIIFTRI
Mendeleevo, Russia
alex.malyshev9@gmail.com

Abstract— A method to calculate the phase center in the near field zone is proposed. It is based on digital processing of the signals received by the probe antenna. The method uses information only about the phase pattern of the probe antenna and the measured phase of the electromagnetic waves. Unlike other methods for calculation the phase center in the near field zone, it has a minimum number of assumptions and approaches and is applicable in practice.

Keywords—antenna measurements, phase pattern, phase center, near field, radio navigation

I. INTRODUCTION

Phase center of electromagnetic radiation is a coordinates where equivalent point source can be placed. In other words, phase center is the origin of equivalent spherical wave [1]. Phase center measurement is important for radio navigation, design and manufacturing of mirror antennas, evaluating proper antenna characteristics and their application on practice, etc.

An arbitrary electromagnetic wave radiated by sources tends to the spherical wave with the increasing distance. These conditions are called far field (FF) approximation. Many important antenna characteristics are defined only in FF and phase center is one of them.

Many theoretical and practical methods to calculate phase center in FF was developed. The most general methods are based on phase measurements. Following algorithms may include least square method (LSM) [2], weighing LSM [3] or iterative technique [4]. Theoretical methods were developed for certain antenna types [5-10]. There is known general theoretical method to calculate phase center based on Poynting vector determination [11]. Its most disadvantage for practice is need to measure all six components of electric and magnetic field. The most common probe antenna in microwave range is open waveguide. Many scientists and engineers still considered an open waveguide detects only tangential to its aperture components of electric field [12]. For this reason, method [11] is applicable in microwave range only with full probe correction algorithms [13-14].

In general case, phase center of spatially distributed sources could not exist. Thus, phase center transforms from a point to

some point cloud. From this point of view decreasing of distance to sources leads to changing of the point cloud. Thus, phase center as a center of point cloud could be determined wherever in near field (NF) zone. Although phase center in NF could be differing from phase center in FF there are some reasons for its measurement. First, various NF microwave systems are under development for communication, radar imaging, medicine and other applications [15-16]. Secondly, error of phase center measurements after NF-FF transformations [13] may be greater than the mismatch between phase center in FF and NF.

There are known methods to calculate phase center in NF. The first is based on the spherical wave expansion of electric field and minimization of significant mode number [17]. However, it is applicable for spherical scanning scheme only. Moreover, determined spherical wave expansion is one for both NF and FF conditions. Therefore, there is no objective discrepancy between the phase center in the NF and FF. The second known method assumes to reconstruct phase pattern from arbitrary measured points to sphere and use methods for FF [18].

All known methods to calculate phase center in NF use assumptions and algorithms valid for FF conditions. It leads to emergence of the additional methodological errors. Method [11] is applicable in NF and does not depend on any FF assumptions. Unknown in microwave range measurements practice Poynting vector may be replaced by wave vector evaluated via phase gradient [19-20]. Thus, a new method to calculate phase center in NF without any FF assumptions may be developed.

Objective of the paper is to improve the accuracy of the phase center measurements in near field zone. To do this in section II we briefly describe phase center theory and possible mathematics solutions. In section III, we represent results of different methods validation over the simulated data. Section IV is dedicated discussion and conclusions.

II. MATHEMATICS FORMULATION

Electric field \mathbf{E} in point \mathbf{r} of the spatially distributed sources in points \mathbf{r}' radiating on a single harmonic is described by the following expression [21]:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{c} \int_V \left(ik\mathbf{G}(\mathbf{r}, \mathbf{r}') \mathbf{j}_e(\mathbf{r}') - \nabla \mathbf{G}(\mathbf{r}, \mathbf{r}') \mathbf{j}_m(\mathbf{r}') \right) dV', \quad (1)$$

where c is the light speed in vacuum, V is the volume containing all sources, k is the wavenumber, \mathbf{G} is the dyadic Green's function, \mathbf{j}_e and \mathbf{j}_m is the electric and magnetic currents densities.

A. Phase Center Existence

Expression (1) is applicable for any point \mathbf{r} mismatching with the sources. Thus, it describes NF conditions. While the distance to sources $\mathbf{R} = \mathbf{r} - \mathbf{r}'$ rises, amplitude factor becomes R^{-1} one for all sources. Expression (1) transforms into the following one:

$$E^i = \frac{ike^{-ikr}}{cr} \int_V \left(\left(\delta_j^i - v^j v_j \right) j_e^j - \varepsilon_{ikj} v^k j_m^j \right) e^{ik(\mathbf{v}, \mathbf{r}')} dV', \quad (2)$$

where $\mathbf{v} = \mathbf{R}/R$ is the sight unit vector.

Expression (2) describes FF approximation. The multiplier outside the integral sign is obviously spherical wave. The problem of phase center existence is in expression under integral sign. If the phase center exists, there is a point \mathbf{r}_c for which phase of the integral is constant for any direction \mathbf{v} . Let us rewrite (2) in form of spherical wave:

$$E^i = f^i(\mathbf{n}) e^{-ik|\mathbf{r} - \mathbf{r}_c|} / |\mathbf{r} - \mathbf{r}_c|, \quad (3)$$

where f^i is the combined amplitude and polarization pattern, \mathbf{n} is the outward unit vector normal to sphere with center in coordinates system origin.

According to expression (3) phase pattern $\Phi(\mathbf{n})$ depends on following expression only:

$$\Phi(\mathbf{n}) = -k|\mathbf{r} - \mathbf{r}_c|. \quad (4)$$

B. Phase Center Calculation in FF

FF conditions include the following approximation:

$$|\mathbf{r} - \mathbf{r}_{PC}| = r \sqrt{1 - 2 \left(\frac{\mathbf{r}}{r}, \frac{\mathbf{r}_c}{r} \right) + \frac{r_c^2}{r^2}} \approx r - (\mathbf{n}, \mathbf{r}_c). \quad (5)$$

On practice, only relative phase measurements are available. Thus, we can construct from (4) and (5) the following system of linear algebraic equations:

$$\begin{pmatrix} n_1^x & n_1^y & n_1^z \\ \dots & \dots & \dots \\ n_N^x & n_N^y & n_N^z \end{pmatrix} \begin{pmatrix} r_c^x \\ r_c^y \\ r_c^z \end{pmatrix} = \frac{1}{k} \begin{pmatrix} \Delta\Phi_1 \\ \dots \\ \Delta\Phi_N \end{pmatrix}. \quad (6)$$

C. Phase Center Calculation in NF

We can assume that (3) is valid also in NF. However, approximation (5) is invalid in these conditions. Let us write the phase gradient (4):

$$\nabla\Phi = -k \frac{\partial}{\partial r^i} |\mathbf{r} - \mathbf{r}_c| = -k \frac{r^i - r_c^i}{|\mathbf{r} - \mathbf{r}_c|} \equiv -\mathbf{k}. \quad (7)$$

Expression (7) means that phase gradient gives wave vector \mathbf{k} . Therefore, we can construct system of linear algebraic equations from (7) with only assumption of phase center existence:

$$\begin{pmatrix} 1 & 0 & 0 & -\frac{\partial\Phi_1}{k\partial x} & \dots & 0 \\ 0 & 1 & 0 & -\frac{\partial\Phi_1}{k\partial y} & \dots & 0 \\ 0 & 0 & 1 & -\frac{\partial\Phi_1}{k\partial z} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 1 & 0 & \dots & -\frac{\partial\Phi_N}{k\partial z} \end{pmatrix} \begin{pmatrix} r_c^x \\ r_c^y \\ r_c^z \\ \rho_1 \\ \dots \\ \rho_N \end{pmatrix} = \begin{pmatrix} r_1^x \\ r_1^y \\ r_1^z \\ \dots \\ r_N^z \end{pmatrix}, \quad (8)$$

where ρ is the distance from observation point \mathbf{r} to its apparent phase center.

Systems (6) and (7) are the core for phase center calculation algorithms. System (6) needs a spherical phase distribution in FF, while (8) is applicable for any distributed measurement points in NF. That is why we choose (8) for calculation phase center in NF.

III. METHODS VALIDATION FOR SINGLE DIPOLE

Comparison of various methods for calculation phase center in NF through experimental data has many problems. First, phase center standard is needed. Direct phase center measurements are unknown. Thus, phase center of the standard might be evaluated via one of the known indirect methods [2-4]. Therefore, we can compare only one method with others. The same situation appears for simulated data comparison except one type of source. Single dipole radiation is described by (2) without integration over the volume. Moreover, FF conditions are met on distance about several wavelengths λ for it. It means we can validate methods on short distance usually corresponding to NF conditions.

A. Numerical Experiments Description

Let us introduce Cartesian coordinate system $Oxyz$ shown on Fig. 1. Place the Hertz's dipole, source described by currents $\mathbf{j}_e = (1 \ 0 \ 0)^T$ and $\mathbf{j}_m = (0 \ 1 \ 0)^T$ in point $\mathbf{r}_c = (0 \ 0 \ z_c)^T$. The most widely spread NF measurement systems have planar scanning scheme. That is why we reproduce NF measurements on the normal to Oz axis plane with center in point $\mathbf{r}_{pln} = (0 \ 0 \ 5\lambda)^T$. Measurement points on the plane are chosen as 61×61 uniform square grid with $\lambda/3$ translations. For methods based on (6) let us introduce spheres in NF with radius $r_{sph} = 5\lambda$ and in FF. Grid for both spheres is from 0° to 90° for polar angle and from -180° to 180° for azimuth angle with 6° translations.

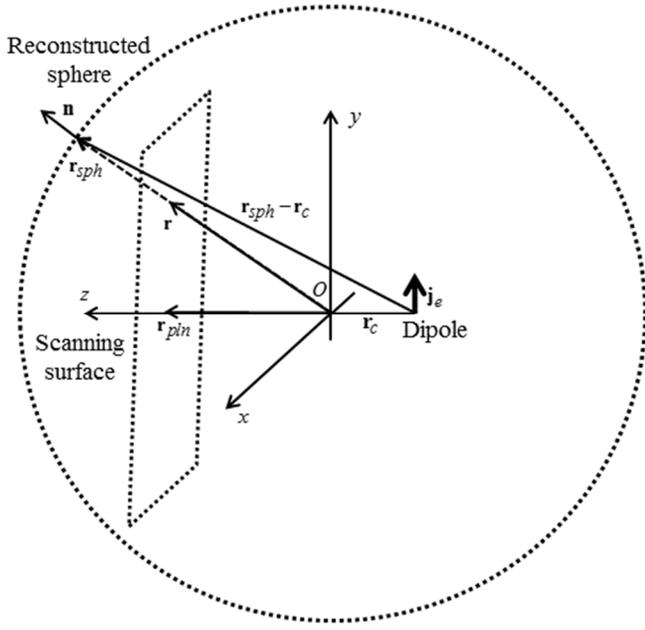


Fig. 1. Scheme of the numerical experiments

B. Brief Methods Description

Now give short description for comparing methods to calculate phase center. Method 1 solves (6) by LSM for FF measurements [2]. Method 2 does the same for electromagnetic field after NF-FF transformation from NF measurements on the plane.

Method 3 uses the same LSM solution of (6) for phase pattern on the sphere in NF, which is evaluated via the following expression [18]:

$$\Phi(\mathbf{r}_{sph}) = \Phi(\mathbf{r}) - k(r_{sph} - r).$$

Method 4 solves (8) for Poynting vectors. Thus, it is close to method [11]. The last original method 5 solves the same (8) for wave vectors evaluated from phase gradient.

C. First numerical experiment

The first numerical experiment is to move Hertz's dipole on Oz axis from -3 to 3λ . Its results for all five methods are in Fig. 2.

Methods 1 and 4 demonstrates the best accuracy. Method 5 accuracy is slightly worse but still very good. Planar NF-FF transformation for Hertz's dipole radiation seems not legit. That is why Method 2 accuracy is the worst. However, it becomes satisfactory for the shortest distance 2λ between Hertz's dipole and scanning plane. Method 3 accuracy rises with z_c magnitude. It is explained by (5) applicability. Large ratio between distances to phase center and measurement points leads to less applicability of (5).

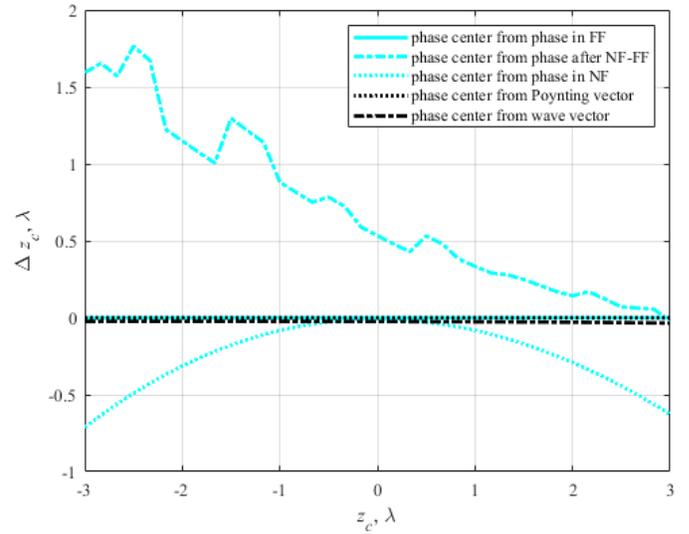


Fig. 2. Phase center calculation error from dipole position

D. Second numerical experiment

The second numerical experiment is to calculate phase center via methods 1, 4 and 5 from spheres with rising radius and constant source coordinate $z_c=3\lambda$. Thus, we verify the methods applicability for NF and FF conditions. Its results are in Fig. 3.

Applicability of (5) rises with sphere radius. Therefore, method 1 error tends to constant value. Method 4 accuracy is close to perfect. Method 5 accuracy decreases from sphere radius. It is explained by rising numerical errors from phase gradient evaluation via finite differences.

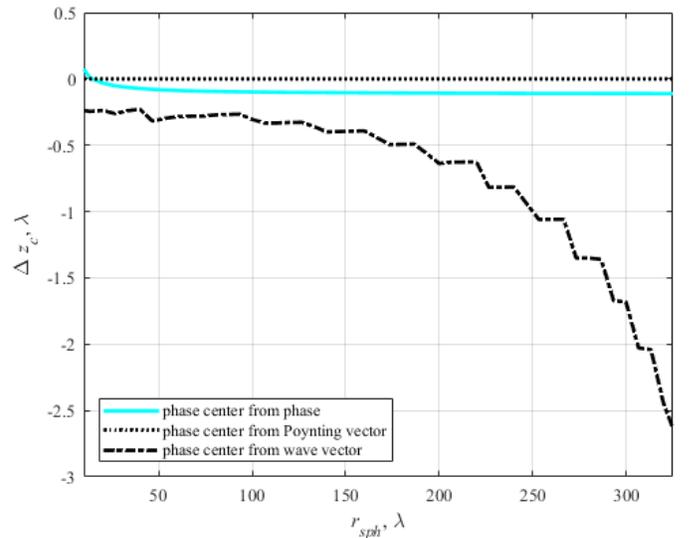


Fig. 3. Phase center calculation error from sphere radius

IV. DISCUSSION AND CONCLUSIONS

Let us summarize results obtained above. There are two pairs of decision conditions. The first is the applicable electromagnetic field approximation, i.e. radiation zone. The second is the type of data: simulated or experimental.

A. Phase Center Calculation Scenarios

If phase center calculation in NF from simulated data is needed method 4 has the best accuracy, i.e. LSM solution of (8) for Poynting vectors. Method 4 is good in FF too if the field is simulated for space coordinates instead angular.

If phase center calculation in FF from simulated data is needed method 1 has the best accuracy, i.e. LSM solution of (6). Nevertheless, there are problems to determine and unwrap phase correctly.

If phase center calculation in NF from experimental data is needed the original method 5 has the best accuracy, i.e. LSM solution of (8) for wave vectors evaluated from phase gradient. However, if phase center in FF from NF measurements is interested the best choice is method 2, i.e. method 1 after NF-FF transformation.

If phase center calculation in FF from experimental data is needed method 1 again has the best accuracy.

B. Conclusions

The main problem of method 4 is lack of data from open waveguide probe for determination Poynting vector in practice. Both electric and magnetic vectors can be evaluated [20]. However, they are evaluated via phase gradient. It means method 4 becomes method 5. The main problem of method 3 is applicability of (5). For NF measurements, method 5 has better performance. The best conditions for method 3 are none spherical FF measurements when (5) is legit.

REFERENCES

- [1] C. A. Balanis, "Antenna theory: analysis and design." 2016.
- [2] A. A. Borodulin, "Opredelenie fazovogo tsentra izluchatelya po metodu naimenshih kvadratov." Raditehnika, 1958, vol. 13, no. 7, pp. 127-150.
- [3] P. N. Betjes, "An Algorithm for Automated Phase Center Determination and its Implementation." AMTA, 2007.
- [4] Y. G. Wang, et al. "A novel method to calculate the phase center of antennas." Journal of Electromagnetic Waves and Applications, 2008, vol. 22, no. 2-3, pp. 239-250.
- [5] M. Teichman, "Precision phase center measurements of horn antennas." IEEE Transactions on Antennas and Propagation, 1970, vol. 18, no. 5, pp. 689-690.
- [6] E. I. Muehldorf, "The phase center of horn antennas." IEEE Transactions on Antennas and Propagation, 1970, vol. 18, no. 6, pp. 753-760.
- [7] K. Rong, et al. "The Research of Phase Center of Log-periodic dipole antenna." Microwave, International Symposium on Antenna, Propagation and EMC Technologies for Wireless Communications, 2007, pp. 735-738.
- [8] J. Tie-hua, et al. "Calculation for the Phase Center of LPDA Based on Simulated Annealing Algorithms and MOM." Environmental Electromagnetics, The 2006 4th Asia-Pacific Conference, 2006, pp. 670-673.
- [9] A. Kumar, et al. "Improved Phase Center Estimation for GNSS Patch Antenna." IEEE Transactions on Antennas and Propagation, 2013, vol. 61, no. 4, pp. 1909-1915.
- [10] I. Vinter, A. Lebedev "Method of determining the phase center of an antenna using polarization measurements." Measurement Techniques, 1998 vol. 41, no. 4, pp. 378-382.
- [11] D. Harke, et al. "A new method to calculate phase center locations for arbitrary antenna systems and scenarios." 2016 IEEE International Symposium on Electromagnetic Compatibility (EMC). IEEE, 2016.
- [12] T. E. Tice, et al. "Probes for microwave near-field measurements." IRE Transactions on Microwave Theory and Techniques, 1955, vol. 3, no. 3, pp. 32-34.
- [13] A. Yaghjian, "An overview of near-field antenna measurements." IEEE Transactions on antennas and propagation, 1986, vol. 34, no. 1, pp. 30-45.
- [14] C. H. Schmidt, M. M. Leibfritz, and T. F. Eibert, "Fully probe-corrected near-field far-field transformation employing plane wave expansion and diagonal translation operators," IEEE Transactions on Antennas and Propagation, 2008, vol. 56.3, pp. 737-746.
- [15] E. G. Larsson, et al. "Massive MIMO for next generation wireless systems." arXiv preprint, 2013, arXiv:1304.6690.
- [16] Zh. Xiaodong, and A. G. Yarovoy, "A sparse aperture MIMO-SAR-based UWB imaging system for concealed weapon detection." IEEE Transactions on Geoscience and Remote Sensing, 2010, vol. 49, no.1, pp. 509-518.
- [17] C. Culotta-López, et al. "Radiation center estimation from near-field data using a direct and an iterative approach." 2017 Antenna Measurement Techniques Association Symposium (AMTA). IEEE, 2017.
- [18] Yu. N. Kalinin, "Measuring of antenna phase center coordinates." Antenny, 2014, no. 14, pp. 54-62. (In Russian)
- [19] N. V. Anyutin, et al. "Correction of the Measured Amplitude-Phase Field Distribution in the Near Field from the Directional Pattern of the Probe." Measurement Techniques, 2018, vol. 61, no. 1, pp. 67-71.
- [20] N. V. Anyutin, et al. "Reconstruction Algorithm of Electromagnetic Field in Case of Elliptic Polarization of Near-Field Probe." 2018 IEEE East-West Design & Test Symposium (EWDTS). IEEE, 2018.
- [21] Ch. Tai, "Dyadic Green functions in electromagnetic theory." 1994.

Method of Calculating the Spare Parts System Availability for Electronic Means

Bogdan Evgenyevich Pankovsky
National Research University. «Higher School of Economics»
Moscow, Russia
0000-0001-8317-9189

Sergey Nikolaevich Poleskiy
National Research University. «Higher School of Economics»
Moscow, Russia
spolessky@hse.ru

Abstract— The aim of research is determine the feasibility of using the methodology for calculating availability and mean delay time to demand fulfilling for spare part by local package of electronic means (EM) with a compound structure of spare parts, tools and accessories(SPTA) kit. Methodology is based on the usage of the analytical model “EM – SPTA kit” with an assessment of the failure level. The main advantages of this technique are possibility to split the main structure of the spare parts system to set of simple, possibility of calculating spare parts sets for several types of replenishment strategies, the simplicity of verification the model for assessing the availability rate and the average delay in satisfying a request for a spare part. Possibility of calculating spare parts sets for several types of replenishment strategies. Simultaneously the technique has a significant drawback, such as methodical technique gives an approximate value of the sufficiency indicators with an calculation error. At the conclusion were proved the ability to use electronic methods for calculating the availability’s rate and the average latency in satisfying the request for the spare part of the spare parts kit system for some replenishment strategies, which also allows to calculate the availability rate for the “EM-SPTA system” model, which is proved to be possible for modern structures of the spare parts system.

Keywords— *electronic means, SPTA, repair, tech support, rate failure, availability factor, mean delay time to demand fulfilling, reliability.*

I. INTRODUCTION

Practice shows that the cost of a system of spare parts, tools and accessories (SPTA) are comparable to the costs of electronic means (EM) themselves, so the challenge is to design an SPTA system that provides a given level of reliability EM at minimal cost. The most popular model for compiling SPTA kits at the moment is the calculation of sufficiency measure, namely, the availability factor and the mean time limit for demand in the pair “SPTA Product Kit”. The approaches to calculating and optimizing the inventory of SPTA kits, given in the sources [1-12], allow us to consider products that are completed with a group or local SPTA kit. If EM in the course of operation involves the use of the SPTA system, then it makes sense to speak not about the reliability index of the EM itself, but about the reliability index of the model “EM - SPTA system”. However, the current practice of designing responsible EMs involves the separate design of EM and the SPTA system assigned to them. Therefore, the sufficiency measure of the SPTA system is introduced, which characterizes the decrease in reliability of the pair “EM - specific SPTA system” compared to the reliability of the pair “EM - endless SPTA system”.

II. FORMULATION OF THE PROBLEM

To ensure high operational reliability, EM is provided with a system to ensure their operability, which includes diagnostic and repair tools, sets of spare elements, means of delivery of spare elements, etc. The SPTA system is the set of all stocks of structural elements included in the EM operational system. A possible shortage of spare elements increases the service substitute, and the limited volume of the SPTA system can significantly affect the value of the reliability index EM, and the limited volume of the SPTA cannot be disregarded in the calculations of reliability. Thus, it is necessary to evaluate the minimum composition of the SPTA system with the given requirements for reliability indicators, which will ensure the operation (repair and maintenance) of the product. The SPTA system, as a rule, should contain an optimal set of spare parts, which are sufficient to replace (restore) the failed EM elements.

A. Sufficiency measure SPTA system

The SPTA system's sufficiency measure according to [2, 4, 5] is mean delay time to demand fulfilling Δ^* for spare part; the delay is caused by the possible absence of the necessary spare element in the SPTA system.

Of all the parameters that determine the reliability of EM, the limitations of the SPTA system affect only the time of their repair. The repair time of EM increases with the absence of the necessary spare element in the SPTA system at the moment when it is needed. Average repair time for EM equipped with a specific SPTA system:

$$\tau = \tau_{\infty} + \Delta^* , \quad (1)$$

Where τ_{∞} – mean repair time of parts with a spare item; Δ^* – mean delay time to demand fulfilling for spare part. The time τ_{∞} does not depend on the SPTA system, it is defined to the stage of its design, Δ^* – which is the first sufficiency measure of the SPTA system, is determined by its functioning parameters and structure. When designing EM, the requirements for their reliability are expressed by setting R_0 of the required value of the reliability index. After the system design has been completed, the calculated values of the function $R(\tau)$ – an EM reliability index depending on the mean repair time of parts, can be considered known, provided that the necessary spare element is always available. Then the requirements for an SPTA system that provides a given EM reliability are expressed by a constraint on it sufficiency measure:

$$\Delta^* \leq \Delta_0 = \tau - \tau_\infty, \quad (2)$$

where τ – the root of the equation $R(\tau) = R_0$; τ_∞ – given the mean substitution time of the failed element EM serviceable replacement. The task of designing an SPTA system comes down to finding such an SPTA system, whose sufficiency measure will be no more than Δ_0 . The availability factor of the SPTA system [4, 9, 13, 14] is called the time-average probability that the SPTA system is not in a state of failure:

$$K^* = \lim_{T \rightarrow \infty} \frac{\int_0^T P^*(t) dt}{T} = \frac{T^*}{T^* + \tau^*}, \quad (3)$$

Where $P^*(t)$ – the probability that at time t SPTA system is not in a fault condition; T^* – mean time between failures SPTA system; τ^* – the average length of a single failure SPTA system (index "*" marked SPTA system indicator). The failure of the SPTA system [10] conditionally refers to such a state of the model as "EM - system SPTA", in which EM completely or partially lost operability due to the failure of one of the EM elements, and the SPTA system cannot provide the necessary spare one. From the definition, it follows that the failure of the SPTA system does not necessarily coincide with the failure to fulfill the requirement for an element, but only with such a failure to fulfill the requirement, which leads to EM downtime.

B. Application of the method for calculating sufficiency measure of SPTA system of electronic means

Consider the application of the technique on the example of a personal computer (PC), for which the following source data is used. Given a product consisting of seven compound part (CP): a) the first difficulty level: 1) keyboard; 2) the mouse; 3) monitor; 4) system unit 4.1 power supply; 4.2 video card; 4.3 hard drive; 4.4 motherboard. b) one CP of the 2nd level - the system unit. The structure of the SPTA system in accordance with the planned operating conditions is as follows (see Fig. 1)

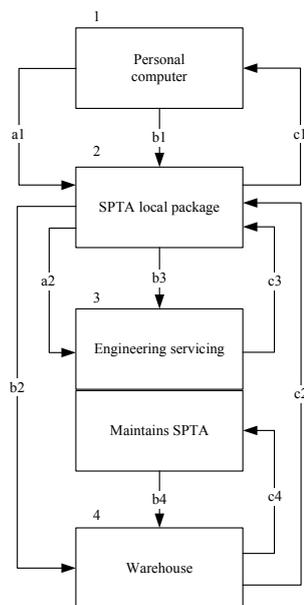


Fig. 1 The structure of the system SPTA PC, combined with the repair part

In the event of failure of the PC element (1), an application (B1) is sent for the supply of a spare one. A defective item itself (a1) is sent along with it. SPTA local package (2) satisfies the request (B1) and assesses at what level the element breaks down (in our case, the product has two levels of difficulty, i.e. it consists of two levels of replacement elements). If an item has failed at level 1, i.e., it cannot be restored by replacing the CP, then an application (b2) to restore (b2) the SPTA local package to the inexhaustible source of replenishment (4) is sent. If the element failed at level 2, i.e., the element failed due to the failure of one of the CP 1 level, then it is sent (a2) to the repair authority (3), where the replacement (b3) (repair, restoration) of the released out of action CP. The application (b3) for replenishment of the SPTA local package (2) is also sent there. Upon exhaustion of the stock of the SPTA repair kit, the repair part (3) submits an application (b4) for restoration (b4) of the SPTA repair kit (3) to an inexhaustible source of replenishment (4).

We introduce the following notation that we will use below:

$N_1 = 4; N_2 = 4$; (N_j – number of component parts in the j -th SPTA kit);

i – part number;

m_i – the number of basic elements of the i -th type in EM;

λ_i – failure rate of the main element of the i -th type in EM;

α_i – type replenishment strategy element of the i -th type EM;

T_{ii} – numeric parameter replenishment strategy (MTTR) of the element i -th type EM;

i_n – the number of spare elements of the i -th type in EM;

$\alpha_i=1$ – periodic replenishment of the failed elements;

$\alpha_i=3$ – repair (restoration) of the failed elements. Any number of elements of this type can be repaired simultaneously;

Λ_{i0} – intensity of demand for the i -type spare part from the SPTA local package;

L – the maximum level of spare parts in the SPTA kit.

The source data for the single and repair kits for the SPTA system are presented in Tables 1 and 2, respectively.

TABLE I. THE SOURCE DATA FOR LOCAL SPTA KIT

Name	i	m_i	λ_i	α_i	T_{ii}	n_i
Monitor	1	2	3,30000	1	8760	2
Mouse	2	1	0,99100	1	8760	1
Keyboard	3	1	2,31000	1	8760	3
System unit	4	1	4,63912	3	336	1

TABLE II. THE SOURCE DATA FOR REPAIR SPTA KIT

Name	i	m_i	λ_i	α_i	T_{ii}	n_i
Motherboard	4.1	1	0,91630	1	8760	1
Video card	4.2	1	0,83392	1	8760	1
HDD	4.3	2	2,49990	1	8760	2
Power Supply	4.4	1	0,38900	1	8760	1

When carrying out the calculation it is taken into account that elements of all types in the PC are not reserved, they are not refused in storage.

To calculate the sufficiency measure structure of the SPTA system (see Fig. 2), we successively calculate the sufficiency measure of the SPTA local package and the SPTA repair kit.

First, we perform calculations for the elements of the SPTA local package.

For elements with periodic replenishment (with $\alpha_i=1$ periodic replenishment) the probability of insufficiency is considered as follows:

$$P_i = \frac{1}{a} e^{-a} \sum_{j=n+2}^{\infty} (j-n-1) \frac{a^j}{j!}; \quad (4)$$

Where $a = m\lambda T_{i1}$.

For a recoverable element (with $\alpha_i=3$ as many as any elements of this type are repaired at the same time), the insufficiency probability is calculated by the formula:

$$P_i = \frac{a^{n+1}}{(n+1)! \sum_{k=0}^{\infty} \frac{a^k}{k!}}; \quad (5)$$

To calculate components 1-3 (see table 1), formula (4) is used, and for component 4, formula (5) is used.

The calculation results are summarized in table 3.

TABLE III. RESULTS OF SUFFICIENCY MEASURE CALCULATION FOR SPTA LOCAL PACKAGE

Part number	a	P
1	0,0290800000	0,00000098927585169
2	0,0086811000	0,00001250604535037
3	0,0202350000	0,00000000137856458
4	0,0015587432	0,00000121294977545

After calculating for the components, we'll go to sufficiency measure for the SPTA kit, which are calculated using the formulas:

$$K_i = 1 - P_{i,l_i+1};$$

$$K = \prod_{i=1}^N (1 - P_{i,l_i+1}); \quad (6)$$

$$\Delta = \frac{1}{\Lambda i 0} \sum_{j=1}^{l_i+1} j P_{ij}$$

The availability factor for the SPTA local package by the formula (6) is equal to

$$K_{LocalKit} = 0,9999852904. \quad (7)$$

Mean delay time to demand fulfilling for spare part by SPTA local package by the formula (6) is obtained equal to

$$\Delta = 1,3086737101 \text{ h.} \quad (8)$$

Similarly, the calculation is carried out kit SPTA repair kit.

Since in our case, the strategy of replenishment and periodic replenishment $\sigma = 0$ (σ – failure rate during storage), the formula has the following form:

$$P_j = \frac{1}{a} e^{-a} \sum_{k=n+j+1}^L \frac{a^k}{k!}. \quad (9)$$

The indicators of the component parts of the SPTA repair kit, calculated using formula (9), are shown in Table 4.

TABLE IV. RESULTS OF SUFFICIENCY MEASURE CALCULATION FOR SPTA REPAIR KIT

Part number	a	P
1	0,0080267880	0,00001067378296054
2	0,0073051392	0,00000884558866041
3	0,0218991240	0,00000042999482805
4	0,0034076400	0,00000193039560329

Sufficiency measure for the SPTA repair kit is calculated using the formulas

$$K = \prod_{i=1}^N (1 - P_{i,l_i+1});$$

$$K_i = 1 - P_{i,l_i+1}; \quad (10)$$

$$\Delta = \frac{1}{\Lambda} \sum_{i=1}^N \sum_{j=1}^{l_i} j P_{ij}$$

Calculate the availability factor of the SPTA repair kit by the formula (10):

$$K_{GroupRepairKit} = 0,9999781203, \quad (11)$$

And also, mean delay time to demand by the SPTA repair kit by the formula (10):

$$\Delta = 4,7163604416 \text{ h.}$$

C. Total availability factor of the EM - SPTA system model

The calculation of the total availability factor of the model "EM - SPTA system" will be carried out using two different methods [15, 16], with the aim of their subsequent comparison.

- 1) According to the methodology from the reference book [16], the following results were obtained:

$$K = K_{EM} \cdot K_{Kit} \cdot K_{RepairKit} =$$

$$= 0,9999174704 \cdot 0,9999852904 \times$$

$$\times 0,9999781204 = 0,9998808840. \quad (13)$$

- 2) Using the ASONIKA-K-ZIP system [15], we obtain the following availability factor:

$$\begin{aligned} K &= K_{EM} \cdot K_{Kit} \cdot K_{RepairKit} = \\ &= 0,9999174704 \cdot 0,9999865033 \times \\ &\times 0,9999779254 = 0,9998819020. \end{aligned} \quad (14)$$

From the calculations we see that the availability factor for the model "EM - system SPTA", obtained using the traditional method [1, 6, 7, 9] and using the presented method for the considered system SPTA [16], are close in meaning. This confirms that the use of the previously proposed method for calculating the described SPTA system (see Fig. 1) is valid and permissible.

CONCLUSION

The method of calculating the availability factor and mean delay time for the SPTA system for some part of the SPTA sets allows us to obtain a quantitative estimate of the availability factor and the average repair time of the model "EM - System SPTA" for arbitrary structures of the SPTA system.

The approach taken from [16], in which the SPTA system first calculates the sufficiency measure, and then corrects the reliability index of the product EM (availability factor or average repair time) with their help: it is possible to decompose the main task; evaluate the sufficiency measure of SPTA kits using different strategies for replenishing the spare elements of the SPTA kits (periodic, periodic with emergency delivery, continuous, according to the level of minimum stock); ease of verification of the calculation model.

A distinctive feature of the developed technique is the direct inclusion of the sufficiency measure of the SPTA system into the EM - System SPTA reliability model.

The obtained results prove the relevance of the work, as well as the effectiveness of the methodology. This technique facilitates the calculation of standard reliability and sufficiency measures, namely the availability factor and the average repair time.

Obviously, the considered method of calculating the model "EM - SPTA system" has limitations: using the methodology for EM that does not have redundancy leads to unjustified redundancy of SPTA kits; Optimizing SPTA kits for just one of the parameters (cost, weight, or volume) leads to difficulties in designing an SPTA kit while simultaneously limiting it with two or more parameters.

Despite the above limitations, the EM-System SPTA model's sufficiency measure calculation method can be applied in the early design stages for an approximate assessment of the composition of the SPTA system.

REFERENCES

- [1] Cherkesov G. N. Otsenka nadezhnosti s uchetom ZIP [Evaluation of reliability with allowance for spare parts]. BHV-Petersburg Publ., 2012, 480 p. (in Russian).
- [2] Military handbook-472 Maintainability prediction. Washington D.C., Department of defense, 1966. 298 p.
- [3] Ushakov I.A. Gnedenko B. V. Probabilistic reliability engineering USA: A Wiley-Interscience Publication, 1995. 663 p. DOI:10.1002/9780470172421
- [4] Ushakov I. A. Kurs teorii nadezhnosti sistem [Course in the theory of system reliability]. Moscow, Drofa Publ., 2008. 239 p. (in Russian).
- [5] Avdeev D. K., Zhadnov V. V. Avtomatizatsiya proektirovaniya sistem zapasnykh chastey i instrumentov [Automation of design of spare parts systems]. Nove informatsionnye tekhnologii i menedzhment kachestva (NIT & QM). Materialy mezhdunarodnogo foruma [New information technologies and quality management (NIT & QM). International forum]. "Quality" Foundation Publ., 2009. pp. 130-133 (in Russian).
- [6] G Pan, Q Luo, X Li, Y Wang. Model of Spare Parts Optimization Based on GA for Equipment. 2018 3rd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2018) pp. 44-47. doi: <https://doi.org/10.2991/msam-18.2018.10>
- [7] Y-k Chen, Q Gao, X Su, S Fang, C Guo. Research on optimization of spare parts inventory policy considering maintenance priority. International Journal of System Assurance Engineering and Management, 2018 vol. 9, pp. 1336-1345
- [8] Mansik H., Burcu B. K. Charles P. S. End-of-life inventory control of aircraft spare parts under performance based logistics. International Journal of Production Economics, 2018, vol. 204, pp. 186-203. doi: <https://doi.org/10.1016/j.ijpe.2018.07.028>
- [9] Epstein J., Ivry O., Spare Parts Supply Chain Shipment Decision Making in a Deterministic Environment. Modern Management Science & Engineering, 2017, vol. 5, no. 1, 10 p.
- [10] Zhadnov V. V., Karapuzov M. A., Kulygin V. N., Poleskiy S. N. Sravnenie lokalnykh vychislitelnykh setey po kriteriyu trebovaniy k komplektam zapasnykh chastey [Comparison of local computer networks by the criterion of requirements to sets of spare parts]. Vestnik komp'yuternykh i informatsionnykh tekhnologii, 2015, no. 4, pp. 36-44. doi: 10.14489/vkit.2015.04.pp.036-044 (in Russian).
- [11] Hekimoğluab M., Van der Laana E., Dekker R. Markov - modulated analysis of a spare parts system with random lead times and disruption risks. European Journal of Operational Research, 2018, vol. 269, pp. 909-922.
- [12] Rahimi-Ghahroodia S., Al Hanbalib A., Vliegenc I.M.H., Cohend M.A. Joint optimization of spare parts inventory and service engineers staffing with full backlogging. International Journal of Production Economics Vol. 212, 2019, pp. 39-50
- [13] Kofanov Iu. N., Zhadnov V. V. Osnovy teorii nadezhnosti i parametricheskoi chuvstvitel'nosti radioelektronnykh sredstv [Basics of the theory of reliability and parametrical sensitivity of radioelectronic devices]. Moscow, Moscow State University of Electronics and Mathematics Publ, 1990. 36 p. (in Russian).
- [14] Spravochnik. Nadejnost' ehlektroradioizdelij [Handbook. Reliability electronic radio elements]. Moscow, 22 Military Scientific Committee of the Armed Forces, 2006. 641 p. (in Russian).
- [15] Poleskiy S. N. Analiz rezul'tatov raschetov nadezhnosti v podsisteme ASONIKA-K [Analysis of the reliability calculation results using ASONIKA-K system]. Tezisy dokladov nauchno-tekhnicheskoi konferentsii studentov, aspirantov i molodykh spetsialistov, posviashchennaia 40-letiiu MIEM [Abstracts of papers of the scientific and technical conference of the students, postgraduates and young specialists, dedicated to the 40th anniversary of MIEM]. Moscow, 2002. pp. 197-198 (in Russian).
- [16] Belyaev Yu. K., Bogatyrev V. A., Bolotin V. V. Nadezhnost tekhnicheskikh sistem [Reliability of technical systems]. Moscow, Radio i svyaz Publ., 1985. 608 p. (in Russian).

Ternary Questionnaires

Dmitrii V. Efanov,
DSc, Professor at “Automation, Remote Control
and Communication on Railway Transport”,
Russian University of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

Valerii V. Khóroshev,
PhD Student, Department
of “Automation, Remote Control
and Communication on Railway Transport”,
Russian University of Transport (MIIT)
Moscow, Russia
Hvv91@icloud.com

Abstract—This paper describes the research results in the questionnaire theory. The authors carried out work on the ternary questionnaire study. These questionnaires include questions that have three outcomes. For solving identification and tracking tasks, in automation devices and computing equipment technical diagnostics, such questionnaires can be more effective than widely used binary questionnaires. These tasks include the technical diagnostics tasks with a time limit for the procedures. The authors describe the questionnaire theory basic concepts, classify them according to various criteria. The ternary questionnaires with various parameters of questions and identifiable events are distinguished in the polychotomous questionnaire class. Methods for optimizing ternary questionnaires are indicated. The authors adapted and described the root question method for optimizing ternary questionnaires. Also, in this paper, the authors focused on questionnaires with faults and with indefinite answers. Ways to create questionnaires by questionnaires, check-lists with faults and with the indefinite answer, as well as features of their optimization are described. Ternary questionnaires form a polychotomous questionnaire class, following the binary questionnaire class, and knowledge of working with them opens the way to exploring other special types of questionnaires.

Keywords—*technical diagnostics; tracing; identification; questionnaire theory; optimization; binary questionnaire; ternary questionnaire; optimal questionnaire.*

I. INTRODUCTION

For solving tracking and discrete search tasks, which include the technical diagnostics tasks of automation devices and computers [1, 2] often use a single mathematical instrument of the questionnaire theory [3, 4].

As applied to the technical diagnostic tasks, the questionnaire theory involves the description of the diagnostic algorithm using questionnaires – tree-oriented weighted graphs [5]. The theory main tasks is to obtain the best questionnaire for a criterion. Such signs may include the event identification average time in the diagnostics algorithm, the procedure implementation average cost, the diagnostic efficiency average value. Among the many questionnaires corresponding to the diagnostic algorithms, it is required to find one or several questionnaires that are optimal according to the chosen criterion.

For solving the questionnaire optimization task using accurate and approximate methods based on the enumeration operations [3, 4]. The optimization method is selected based on the questionnaire structure analysis and the features of the nodes. Vertices correspond to checks and identifiable events in the diagnostic algorithm. Questionnaires may be different in their appearance trees. Binary questionnaires are widely used – in such questionnaires, each question has two out-

comes corresponding to two different answers “yes” and “no” (“1” and “0”, “identified” and “not identified”) [6 – 9]. Binary questionnaires are the main and simplest questionnaire type. For some tasks, other questionnaire types are used, where the outcomes are interpreted with many events [10, 11].

As is known in [3, 4], any questionnaire can be easily converted to binary. Such a conversion will lead to an increase in the number of questions in relation to the original questionnaire. For some tasks, this is not critical, and for some, it has a negative effect. For example, if solving a technical diagnostics tasks, there is a limit on the procedure maximum duration, a non-binary questionnaire may turn out to be a more effective than binary.

This paper is devoted to presenting the results of a study of this questionnaire type, in which each question has three outcomes. Such questionnaires are named by the authors by ternary questionnaires. In addition, the authors consider specific types of questionnaires, such as questionnaires with faults and with indefinite answers (with don't cares).

II. QUESTIONNAIRES: REPRESENTATION, CLASSIFICATION, OPTIMIZATION

As mentioned earlier, the questionnaire is a tree-oriented graph. There are no cycles in this graph. The questionnaire type is determined by the peculiarities of its nodes. Nodes in the questionnaire can have outgoing arcs, only an incoming arc (in the tree graph it is always the same) or both. The first and third types of nodes correspond to the root question and intermediate questions, and the second to the identified events. The questions set $Y = \{y_1, y_2, \dots, y_n\}$ is intended to divide among themselves the identifiable events sets $X = \{x_1, x_2, \dots, x_m\}$. Each question $y_i \in Y$ divides the events sets X into some non-intersecting subsets — answers or outcomes of question y_i . The number of such subsets is determined by the base $\alpha(y_i)$ of the question $y_i \in Y$ – the number of arcs emanating from the corresponding node. Questionnaires can also be given in the matrix form like the questionnaire check-list [6, 7], in which the identified events $X = \{x_1, x_2, \dots, x_m\}$, are placed by columns, and the available questions $Y = \{y_1, y_2, \dots, y_n\}$ are arranged in rows to solve the separation tasks. At the row and column, intersection indicates the event belongs to one or another outcome of the question.

If all the questions in the questionnaire have the same number of outcomes, then such a questionnaire is called homogeneous. If the base of at least one question differs from the base of the remaining questions, then the questionnaire is heterogeneous [12]. Homogeneous questionnaires can be divided into classes that correspond to the same grounds:

binary questionnaires, for which $\forall y_i \in Y \alpha(y_i) = 2$ (they are also called dichotomous questionnaires), and polychotomous questionnaires, for which $\forall y_i \in Y \alpha(y_i) > 2$.

Definition 1. The ternary questionnaires are the questionnaires for which:

$$\forall y_i \in Y \alpha(y_i) = 3. \quad (1)$$

Binary questionnaires and methods for their optimization have been studied quite well and are widely used in practical tasks [6 – 9]. Ternary questionnaires are the next most complex type of polychotomous questionnaires.

The nodes in the questionnaire are weighted, and some weight functions $\omega(y_i)$ and $\omega(x_j)$, correspond to them, where $y_i \in Y$ and $x_j \in X$. If the weight function value is constant, then we can assume that it is a question weight coefficient. The weight coefficients values are determined and fixed, which allows them to be normalized relative to the identified events set or given by probabilities of an event:

$$p(x_j) = \frac{\omega(x_j)}{\sum_{j=1}^m \omega(x_j)}. \quad (2)$$

Where $\sum_{j=1}^m p(x_j) = 1$.

For each question, the weighting coefficient value is the outcomes weights values sum.

Certain $c(y_i)$ $y_i \in Y$ numbers are assigned to all questions, both root and intermediate nodes of the questionnaire, corresponding to their absolute or relative value. The issue cost in the technical diagnostic tasks can correspond to the test time, the implementation costs, the test final effectiveness, etc.

For each questionnaire, the average cost the identifying event sets $X = \{x_1, x_2, \dots, x_m\}$, called the questionnaire implementation cost, can be determined:

$$C = \sum_{i=1}^n p(y_i) c(y_i). \quad (3)$$

Different questionnaires may have different meanings of the implementation cost. The questionnaire theory main task is to find the questionnaire optimal implementation cost value by criterion C , for which $C = C_{min}$. Methods for optimizing questionnaires are determined based on the questionnaire type and are described in [3, 4, 7]. Such methods include such popular optimization methods as “The branch and bound method” [13] and “The dynamic programming method” [14].

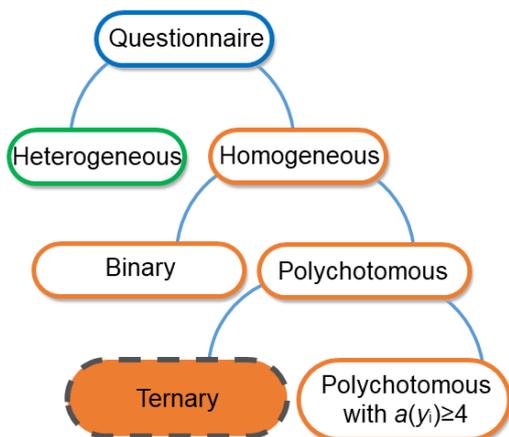


Fig. 1. The questionnaires classification.

For solving the events identification tasks $X = \{x_1, x_2, \dots, x_m\}$ are minimally needed the number of Y_{min} questions. The number of Y_{min} is determined by the question bases. For example, for binary questionnaires, this number is equal to $Y_{min} = \lceil \log_2 m \rceil_{min}$, and for ternary questionnaires it is $Y_{min} = \lceil \log_3 m \rceil_{min}$ then m – the number of identifiable events, [...] – it means integer on top of the calculated value. Thus, the non-redundant question sets Y is set as $|Y| \geq Y_{min}$. If $|Y| = Y_{min}$, the questionnaire is called compact, otherwise it is non-compact.

In conclusion, of the questionnaires, types review and their components are present the questionnaires classification in Fig. 1. Consider further the ternary questionnaires features and their description and optimization.

III. TERNARY QUESTIONNAIRES AND METHODS FOR THEIR OPTIMIZATION

Each question in the ternary questionnaire has three outcomes. Conventionally, we denote them by numbers of outcomes 2, 1, 0, numbering them on the graph from left to right. We assign the normalized weighting factors near each node in the numbers form, and each question implementation cost will be indicated by a number in brackets. In Fig. 2 presents an elementary ternary questionnaire consisting of one question y_1 , which allows dividing the set $X = \{x_1, x_2, x_3\}$. The question implementation cost is $c(y_1)$, the weight is $p(y_1) = p(x_1) + p(x_2) + p(x_3)$.

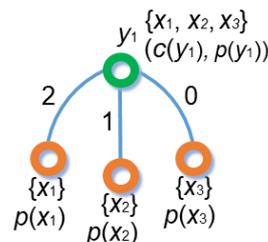


Fig. 2. The questionnaire elements designations.

Any ternary questionnaire consists of an elementary questionnaire set. The one root node of the elementary questionnaires is the questionnaire root and the all others root nodes are the new questionnaire intermediate nodes. Nodes in the questionnaire are ranked relative to each other. The node rank r is a number characterizing its position relative to the nodes corresponding to the root question, equal to the number of arcs leading from the questionnaire root node to a given node.

In Fig. 3 presents the ternary questionnaires examples. For each, both questionnaires node, the implementation cost and weights are indicated.

Using the formula (3), we determine the implementation cost for each questionnaire.

For the first questionnaire (Fig. 3, a) we have:

$$C_1 = 1 \cdot 1.00 + 2 \cdot 0.75 + 2 \cdot 0.35 + 3 \cdot 0.35 = 4.25.$$

For the second questionnaire (Fig. 3, b) we have:

$$C_2 = 1 \cdot 1.00 + 2 \cdot 0.35 + 2 \cdot 0.4 + 3 \cdot 0.25 = 3.25.$$

Comparing each questionnaire implementation cost values, we note that $C_1 > C_2$. Thus, the diagnostic algorithm implemented on the first questionnaire allows identifying

events faster than the algorithm implemented in the second questionnaire.

The examples given are abstract. They show questionnaires special cases. In reality, the number of questions needed to solve the fully identifying events tasks of the set $X = \{x_1, x_2, \dots, x_m\}$ may belong to the set $n \in \{\lceil \log_3 m \rceil, \lceil \log_3 m \rceil + 1, \dots, 3^m\}$. However, for different events subsets received at each new partition stage, some questions may not be effective, that is, not separate events in a subset, have two or three outcomes.

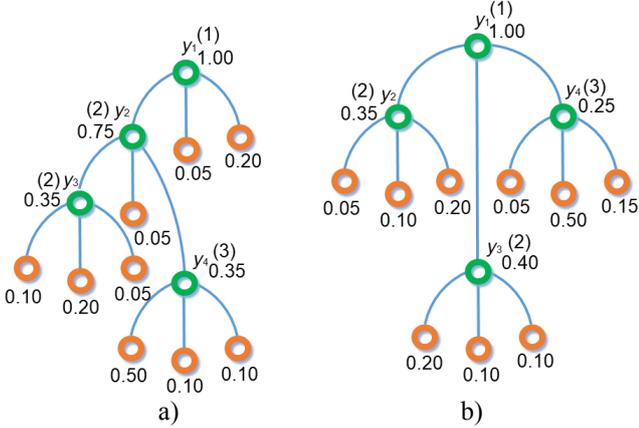
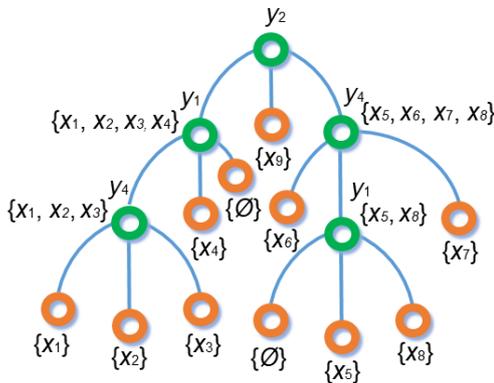


Fig. 3. Ternary questionnaires.

Definition 2. A fictitious outcome is the question outcome with a base $\alpha(y_i) = 3$, which, when installed, decreases the original base value of the question.

The fictitious outcome will be denoted with the sign of the empty set $\{\emptyset\}$. A zero-weight value is assigned to its node.

In reality, a questionnaire that has questions with fictitious outcomes ceases to be ternary and becomes heterogeneous, including questions with bases $\alpha(y_i) = 2$ and $\alpha(y_i) = 3$. The introduction of fictitious outcomes will be required later in solving the questionnaire optimization tasks, and the “addition” of each question to the question with a basis $\alpha(y_i) = 3$ allows you to save the questionnaire type.



For example, let the questionnaire contain four questions $Y = \{y_1, y_2, y_3, y_4\}$, intended to divide the nine events $X = \{x_1, x_2, \dots, x_9\}$. The questions $y_i \in Y$ allow you to divide the original set X into the following subsets:

$$y_1 = \{x_1, x_2, x_3\} \cup \{x_4, x_5, x_6\} \cup \{x_7, x_8, x_9\},$$

$$y_2 = \{x_1, x_2, x_3, x_4\} \cup \{x_5, x_6, x_7, x_8\} \cup \{x_9\},$$

$$y_3 = \{x_1, x_6, x_7\} \cup \{x_2, x_8, x_9\} \cup \{x_3, x_4, x_5\},$$

$$y_4 = \{x_1, x_4, x_7\} \cup \{x_2, x_5, x_8\} \cup \{x_3, x_6, x_9\}.$$

We will not ask specific the identified events implementation costs and weights values. Let's set the questionnaire in the questionnaire check-list form (Tabl. 1).

TABLE I. THE QUESTIONNAIRE CHECK-LIST

y_i	$c(y_i)$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
y_1	$c(y_1)$	2	2	2	1	1	1	0	0	0
y_2	$c(y_2)$	2	2	2	2	1	1	1	1	0
y_3	$c(y_3)$	2	1	0	0	0	2	2	1	1
y_4	$c(y_4)$	2	1	0	2	1	0	2	1	0
$p(x_i)$		$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_4)$	$p(x_5)$	$p(x_6)$	$p(x_7)$	$p(x_8)$	$p(x_9)$

Each question has three outcomes, but only the root question cannot contain a fictitious outcome. All others can be questions with two real and one fictitious outcome. For example, in Fig. 4 presents two options for questionnaires. The first questionnaire (Fig. 4, a) has a design with a node located on three ranks. When asking questions on some identified events subsets there are fictitious outcomes. The questionnaire belongs to the type of the non-compact question since for this question the minimum required several questions for the all events separation is $Y = \lceil \log_3 9 \rceil_{min}$. The second questionnaire (Fig. 4, b) is compact and has no fictitious outcomes.

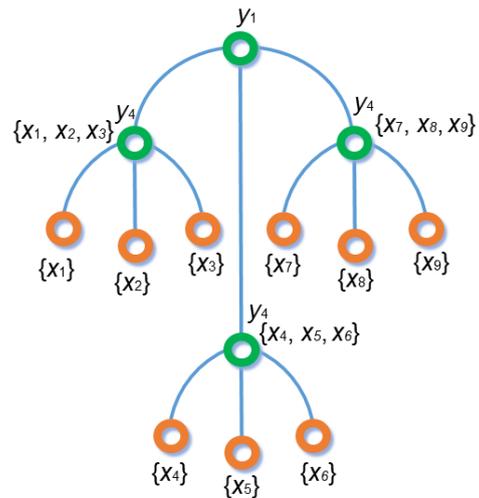


Fig. 4. Samples questionnaire.

The questionnaires theory main task is the questionnaires optimization by the minimum implementation cost criterion. There are various optimization methods that allow obtaining accurate results (optimal questionnaires) and approximate to optimal ones (quasi-optimal questionnaires). The well-known exact methods are “the Branch and Bound method” and “The Dynamic programming method”, as applied to any questionnaire type, described in detail in [3, 4]. These algorithms have exponential complexity and work with a limited number of identifiable events (as a rule, up to 30 ... 35, as they are brute force methods). Approximate methods make it possible to obtain questionnaires with an implementation cost that is close to the minimum [7]. Consider one of these methods – the root question method, described and studied in [6, 7] for binary questionnaires.

IV. THE ROOT QUESTION METHOD FOR TERNARY QUESTIONNAIRES

Any optimization methods associated with the procedure for checking the cost of issues. The following theorem is known, which characterizes the optimal questionnaire [3, 4].

Theorem 1. The optimal questionnaire consists of optimal sub-questions.

According to the statement, the optimal questionnaire can be constructed by combining the optimal sub-questionnaires at each rank.

Adapt the root question method for use with ternary questionnaires. This method involves choosing the first question from a question set, root question, based on a reasonable analysis of all questions outcomes. The next stage is the sequential consideration of the sub-questionnaires obtained for each outcome of the root question. Each new sub-questionnaire is treated as a separate questionnaire and a root question is selected for it. If the task is to identify events completely, then the root question selection procedures end when all events are divided.

The root question method is based on choosing the best question based on the value of the preference function.

A. Relationship comparison between questions

The root question method means comparing the questions among themselves in terms of the outcomes of subsets ratios.

Definition 3. A comparison relationship is established between two questions if for any one of the outcome there is an outcome of the opposite question, the events subset of which is completely included in the events subset of the question under consideration.

For binary questionnaires, the comparison relationship may not always be established, the same applies to homogeneous questionnaires with the bases of all questions $\alpha(y_i) > 2$.

Consider the comparison relationship possible variants between questions in ternary questionnaires.

To do this, follow these steps. The identified events set $X = \{x_1, x_2, \dots, x_m\}$ is comparable to a one-dimensional vector with length m . The question A splits the events set into three subsets $X_A^2 \subset X, X_A^1 \subset X, X_A^0 \subset X$, belonging to each of its outcomes. The question B divides the events set into three subsets $X_B^2 \subset X, X_B^1 \subset X, X_B^0 \subset X$. Each of the outcomes subsets of questions A and B also associate one-

dimensional vectors with lengths $m_{A,2}, m_{A,1}, m_{A,0}$ and $m_{B,2}, m_{B,1}, m_{B,0}$. Since the subsets of the outcomes of each question are disjoint, $m_{A,2} + m_{A,1} + m_{A,0} = m_{B,2} + m_{B,1} + m_{B,0} = m$. Let questions A and B are in relation to the comparison, that is, they satisfy the conditions of *Definition 3*. Let us establish what these relations can be.

Since we have assumed that questions A and B are in relation to comparison, the subsets $X_A^2 \subset X, X_A^1 \subset X, X_A^0 \subset X$ should separately have intersections with the subsets $X_B^2 \subset X, X_B^1 \subset X, X_B^0 \subset X$, which completely include certain events corresponding to a particular outcome. Since the questions are different, all the subsets of their outcomes cannot be equal. For simplicity, we suppose that the comparison relations are established by the same outcome of both questions and the subsets X_A^2 and X_B^2, X_A^1 and X_B^1, X_A^0 and X_B^0 are compared to each other. Thus, $m_{A,2} \neq m_{B,2}, m_{A,1} \neq m_{B,1}$ and $m_{A,0} \neq m_{B,0}$.

Consider the case when one of the considered pairs is the same, for example, a subsets pair X_A^2 and X_B^2 . Then $m_{A,2} = m_{B,2}$. It follows that $m_{A,1} + m_{A,0} = m - m_{A,2}$ and $m_{B,1} + m_{B,0} = m - m_{B,2}$, or that $m_{A,1} + m_{A,0} = m_{B,1} + m_{B,0}$. It follows that between the like outcomes two types of relationships can be established in which: 1) $m_{A,1} > m_{B,1}$ and $m_{A,0} < m_{B,0}$; 2) $m_{A,1} < m_{B,1}$ and $m_{A,0} > m_{B,0}$. They are depicted in Fig. 5. There are no other cases where the same-named outcomes of questions give the same subsets of identifiable events, does not exist.

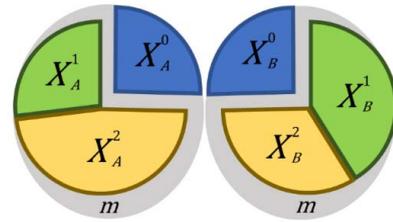


Fig. 5. Comparison relations for questions with equal subsets of similar outcomes.

Consider the case when there are no equal subsets for the outcomes of the same name. In this case, when $m_{A,2} > m_{B,2}$ and $m_{A,2} < m_{B,2}$ there are several variants of relations: 1) $m_{A,1} > m_{B,1}$ and $m_{A,0} < m_{B,0}$; 2) $m_{A,1} < m_{B,1}$ and $m_{A,0} > m_{B,0}$; 3) $m_{A,1} > m_{B,1}$ and $m_{A,0} > m_{B,0}$; 4) $m_{A,1} < m_{B,1}$ and $m_{A,0} < m_{B,0}$. This case is depicted in Fig. 6.

There is another, third, relations variant in which the one subset of the question A outcomes includes the two outcomes subsets of question B , and the other question A two outcomes are fully included in the third question B outcome (Fig. 7). This case is described by comparing the lengths of the vector $m_{A,2} = m_{B,2} + m_{B,1}, m_{B,0} = m_{A,1} + m_{A,0}$.

The analysis shows that there are no other correlations variants between the outcome’s subsets of two comparable questions. Questions that cannot establish comparison relationships are incomparable.

Other relationships variants between ternary questions outcomes subsets are possible, but all of them can easily be reduced to the cases considered.

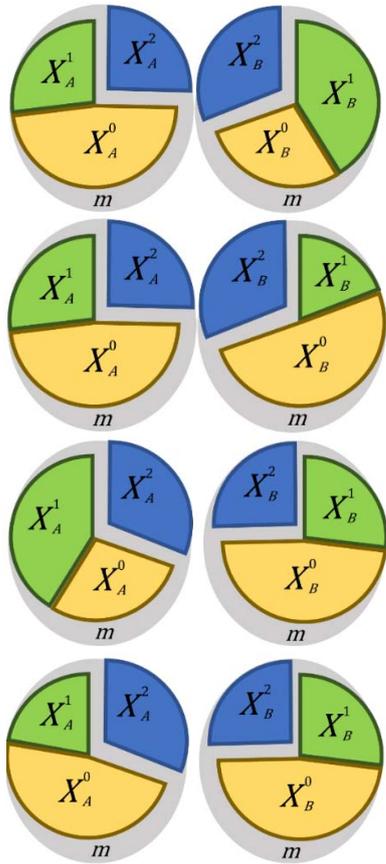


Fig. 6. Comparison relations for questions with unequal subsets of similar outcomes

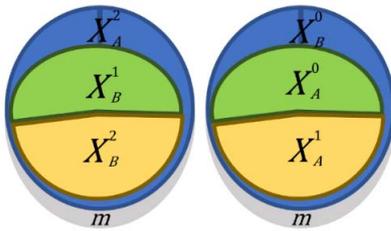


Fig. 7. Comparison relations for questions with fully included events of two outcomes in one outcome of the opposite question.

In Fig. 8 are examples illustrating the various comparison relationships between two ternary questions.

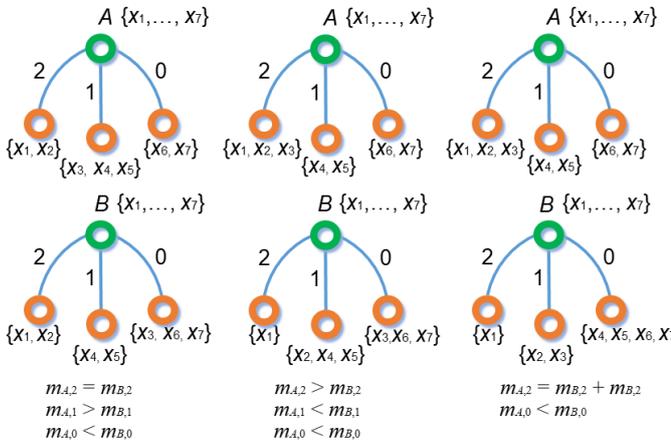


Fig. 8. Questions that are in a comparison relationship.

B. Preference function

If the questions in the questionnaire are comparable, then you can specify the most preferred option between the questions being compared. Preference is determined by considering the events weight and the implementation costs.

Definition 4. Question A is preferable to question B ($A < B$), if, when asking the first question A , and the second question B , the final questionnaire implementation cost will be less than in the opposite case.

To assess the preference of one question over another in the questions pair A and B , the preference function Φ is introduced. This function is determined by considering the implementation costs and the weight of the events for each of the questions outcomes in the ratios defined between them. Let define these relations.

Let question A to be selected as the first question. It divides the full events set $X = \{x_1, x_2, \dots, x_m\}$ into three disjoint subsets $X_A^2 \subset X$, $X_A^1 \subset X$, $X_A^0 \subset X$. Question B is asked second and breaks the full events set $X = \{x_1, x_2, \dots, x_m\}$ into three disjoint subsets $X_B^2 \subset X$, $X_B^1 \subset X$, $X_B^0 \subset X$. This questions pair in relation to comparison, presented in Fig. 9.

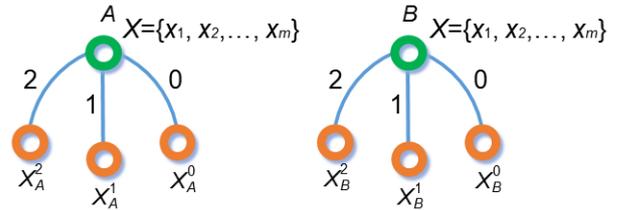


Fig. 9. Question A and B .

Consider the various relations of comparison between the ternary questions and establish a number of regularities for them. These patterns are inherent in all other comparisons variants of questions, taking into account the features of each of the outcomes subsets of each question.

Consider the case when any two outcomes subsets of different questions are equal. For example, X_A^2 and X_B^2 . Two previously identified relationships versions between subsets of other question outcomes appear. These cases correspond to the relations between the lengths of the one-dimensional vectors corresponding to the outcomes subsets:

- 1) $m_{A,1} > m_{B,1}$ and $m_{A,0} < m_{B,0}$;
- 2) $m_{A,1} < m_{B,1}$ and $m_{A,0} > m_{B,0}$.

Let $m_{A,1} > m_{B,1}$ and $m_{A,0} < m_{B,0}$. If question A is asked first, then posing question B on the events set off the third outcome is meaningless (Fig. 10).

This is explained by the following facts. If $X_A^0 \supset X_B^0$, then posing question B will not give a new split of events on a given events subset. Thus, the only effective way to further "crush" the identified events subsets is to pose question B on a events subset X_A^1 . In this case, since $X_A^1 \subset X_B^1$, posing question B will split the subset X_A^1 into three subsets: \emptyset , X_B^1 , $X_A^1 \setminus X_B^1$. Note that the first outcome subset when asking question B after question A gives an empty subset of identifiable events. This is a bogus outcome.

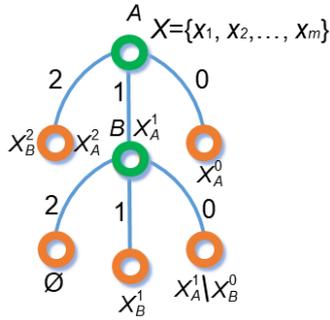


Fig. 10. Sequential formulation of questions A and B .

The opposite way of asking questions is considered in a similar way: first, question B , and then question A (Fig. 11).

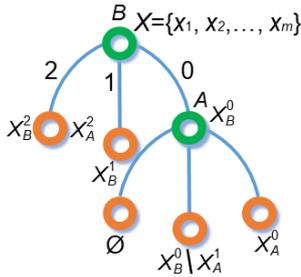


Fig. 11. Sequential formulation of questions B and A .

If $A < B$, then the first questionnaire implementation cost (Fig. 10) should be less than the second questionnaire implementation cost (Fig. 11).

Using the expression (3), we determine the first and second questionnaires implementation costs. In this case, the question implementation cost for splitting the subsets of each outcome will be denoted as $C(X_A^k)$ and $C(X_B^k)$, where $k \in \{0,1,2\}$.

For the first questionnaire we get:

$$C_{AB} = 1 \cdot c(A) + c(X_A^2) + c(X_A^0) + c(X_B^1) + c(X_A^1 \setminus X_B^0) + c(B) \sum_{p_j \in X_A^1} p_j, \quad (4)$$

where $c(A)$, $c(B)$ – A и B implementation costs; $C(X_A^k)$ and $C(X_B^k)$, $k \in \{0,1,2\}$ – X_A^k и X_B^k sub-questionnaires implementation costs.

For the second questionnaire we get:

Questionnaire implementation cost Q_{BA} calculated:

$$C_{BA} = 1 \cdot c(B) + c(X_B^2) + c(X_B^0) + c(X_A^1) + c(X_B^0 \setminus X_A^1) + c(A) \sum_{p_j \in X_B^0} p_j. \quad (5)$$

Determine the costs difference C_{AB} and C_{BA} :

$$\begin{aligned} \Delta C &= C_{AB} - C_{BA} = \\ &= \left(1 \cdot c(A) + c(X_A^2) + c(X_A^0) + c(X_B^1) + c(X_A^1 \setminus X_B^0) + c(B) \sum_{p_j \in X_A^1} p_j \right) - \\ &- \left(1 \cdot c(B) + c(X_B^2) + c(X_B^0) + c(X_A^1) + c(X_B^0 \setminus X_A^1) + c(A) \sum_{p_j \in X_B^0} p_j \right) = \\ &= c(A) - c(B) + c(X_A^1 \setminus X_B^0) - c(X_B^0 \setminus X_A^1) + \\ &+ c(B) \sum_{p_j \in X_A^1} p_j - c(A) \sum_{p_j \in X_B^0} p_j. \quad (6) \end{aligned}$$

Expression (6) can be simplified, since $X_A^1 \setminus X_B^0 = X_B^0 \setminus X_A^1$ (As follows from the considered case ratios subsets $m_{A,2} = m_{B,2}$, $m_{A,1} > m_{B,1}$ и $m_{A,0} < m_{B,0}$).

Then

$$\begin{aligned} \Delta C &= C_{AB} - C_{BA} = \\ &= c(A) - c(B) + c(B) \sum_{p_j \in X_A^1} p_j - c(A) \sum_{p_j \in X_B^0} p_j. \quad (7) \end{aligned}$$

From (7) it follows that the inequality $\Delta C < 0$, which characterizes the preference of question A over question B , is fulfilled under the condition:

$$c(A) + c(B) \sum_{p_j \in X_A^1} p_j < c(B) + c(A) \sum_{p_j \in X_B^0} p_j. \quad (8)$$

Introduce the preference function:

$$\Phi(A, B) = \frac{c(A) + c(B) \sum_{p_j \in X_A^1} p_j}{c(B) + c(A) \sum_{p_j \in X_B^0} p_j}. \quad (9)$$

If $\Phi(A, B) < 1$, question A is preferable, otherwise question B . If $\Phi(A, B) = 1$, the questions are equivalent.

Example 1. Determine the preference function value for the above case (Fig. 10 and Fig. 11), taking into account the questions implementation costs and weights: $c(A) = 2.5$; $c(B) = 2.3$; $p(x_1) = 0.12$; $p(x_2) = 0.1$; $p(x_3) = 0.15$; $p(x_4) = 0.19$; $p(x_5) = 0.11$; $p(x_6) = 0.16$; $p(x_7) = 0.17$.

In Fig. 12 illustrates Example 1 conditions.

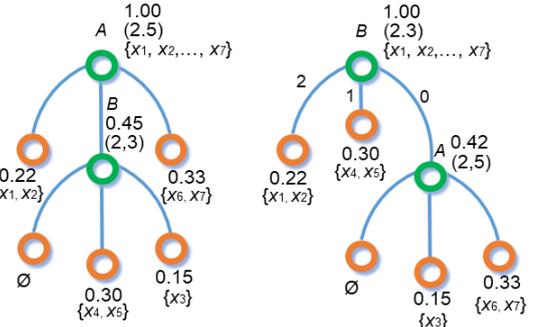


Fig. 12. Questionnaires example 1.

The preference function value, defined by the formula (9) is equal to:

$$\Phi(A, B) = \frac{2.5 + 2.3 \cdot 0.45}{2.3 + 2.5 \cdot 0.48} = \frac{3.535}{3.5} = 1.01.$$

$\Phi(A, B) > 1$, which means that $B < A$.

By analogy with the above reasoning, expressions are derived that characterize the preference ratios for the second and third cases of the relationship between the questions.

Consider these cases and the rationale for the choice of expressions for calculating the functions of preference for them by examples.

Example 2. Define an expression describing the coefficient of preference for two questions A and B : $X_A^2 = \{x_1, x_2\}$, $X_A^1 = \{x_3, x_4, x_5, x_6\}$, $X_A^0 = \{x_7, x_8\}$, $X_B^2 = \{x_1, x_2, x_3\}$, $X_B^1 = \{x_4, x_5\}$, $X_B^0 = \{x_6, x_7, x_8\}$.

Solve the tasks in general.

We consider the second case of the comparison relationship because $X_A^2 \subset X_B^2$, $X_A^1 \supset X_B^1$, $X_A^0 \subset X_B^0$. The illustration for the example in question is shown in Fig. 13.

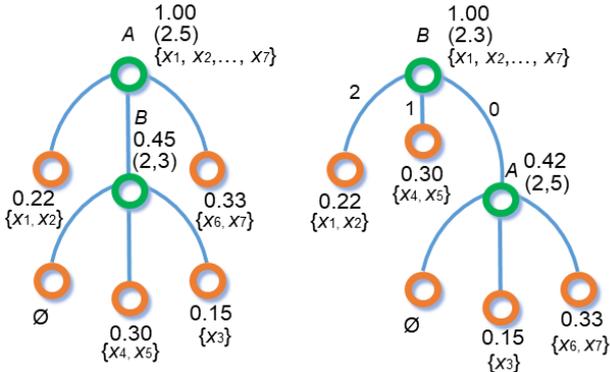


Fig. 13. Questionnaires example 2.

The first questionnaire implementation cost is calculated as:

$$C_{AB} = 1 \cdot c(A) + c(X_A^2) + c(X_A^0) + c(X_A^1 \setminus X_B^2) + c(X_B^1) + c(X_A^1 \setminus X_B^0) + c(B) \sum_{p_j \in X_A^1} p_j. \quad (10)$$

The second questionnaire implementation cost is calculated as:

$$C_{BA} = 1 \cdot c(B) + c(X_A^2) + c(X_A^0) + c(X_B^2 \setminus X_A^1) + c(X_B^1) + c(X_B^0 \setminus X_A^1) + c(A) \sum_{p_j \in (X_B^2 \cup X_B^0)} p_j. \quad (11)$$

From (10) and (11) follows:

$$\Delta C = C_{AB} - C_{BA} = c(A) + c(B) \sum_{p_j \in X_A^1} p_j - (c(B) + c(A) \sum_{p_j \in (X_B^2 \cup X_B^0)} p_j). \quad (12)$$

The preference function, in this case, is written as:

$$\Phi(A, B) = \frac{c(A) + c(B) \sum_{p_j \in X_A^1} p_j}{c(B) + c(A) \sum_{p_j \in (X_B^2 \cup X_B^0)} p_j}. \quad (13)$$

In the numerator and denominator of expression (13), events weights are summed by those outcomes of questions, the subsets of which include the subsets of the corresponding outcomes of the compared question: $X_A^2 \subset X_B^2$, $X_A^1 \supset X_B^1$, $X_A^0 \subset X_B^0$.

Example 3. Find a general expression for calculating the preference function for the third type of comparison relations using the example of two questions A и B: $X_A^2 = \{x_1, x_2, x_3\}$, $X_A^1 = \{x_4, x_5, x_6\}$, $X_A^0 = \{x_7, x_8\}$, $X_B^2 = \{x_1, x_2\}$, $X_B^1 = \{x_3\}$, $X_B^0 = \{x_4, x_5, x_6, x_7, x_8\}$.

From the partitioning classes, obviously, that: $X_A^2 = X_B^2 \cup X_B^1$, and $X_B^0 = X_A^1 \cup X_A^0$. In Fig. 14 depicts relevant options for asking questions.

Writing down the implementation cost both questionnaires Fig. 14, we get:

$$C_{AB} = 1 \cdot c(A) + c(X_A^1) + c(X_A^0) + c(X_B^2) + c(X_B^1) + c(B) \sum_{p_j \in X_A^2} p_j. \quad (14)$$

$$C_{BA} = 1 \cdot c(B) + c(X_B^2) + c(X_B^1) + c(X_A^1) + c(X_A^0) + c(A) \sum_{p_j \in X_B^0} p_j. \quad (15)$$

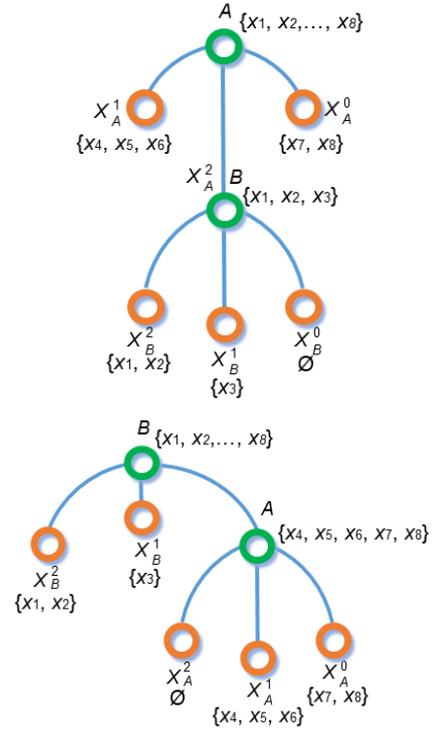


Fig. 14. Questionnaires example 3.

Subtracting (14) from (15), we get:

$$\Delta C = C_{AB} - C_{BA} = c(A) + c(B) \sum_{p_j \in X_A^2} p_j - (c(B) + c(A) \sum_{p_j \in X_B^0} p_j). \quad (16)$$

Where the preference function is determined by the expression:

$$\Phi(A, B) = \frac{c(A) + c(B) \sum_{p_j \in X_A^2} p_j}{c(B) + c(A) \sum_{p_j \in X_B^0} p_j}. \quad (17)$$

In the numerator and denominator of expression (17), there are sums of event weights by those outcomes of questions, the subsets of which contain the subsets of the corresponding outcomes of the compared question. This is similar to expressions (9) and (13).

Based on the above reasoning, one can speak of the unified nature of formulas (9), (13) and (17).

Theorem 2. The preference function value for two questions A and B, regardless of the comparison relationship type established, is determined by the formula:

$$\Phi(A, B) = \frac{c(A) + c(B) \sum_{p_j \in X(a_m)} p_j}{c(B) + c(A) \sum_{p_i \in X(b_k)} p_i}, \quad (18)$$

where $c(A)$, $c(B)$ – are the questions A and B implementation costs, respectively; $\sum_{p_j \in X(a_m)} p_j$ – is the sum of the identified events weights that are included in the subsets of the question A outcomes, which fully include the compared question B events subsets; $\sum_{p_i \in X(b_k)} p_i$ – similar to the previous one.

C. The root question selection

To select the root question, we need to compare $C_n^2 = \frac{n^2-n}{2}$ questions with each other in pairs. In addition, it is necessary to calculate the preference functions values. Moreover, to select the best question, it is convenient to build a special preference graph.

Definition 5. A questionnaire preference graph is a directed graph in which vertices correspond to questions, and the directions of arcs indicate the most preferred question from any questions pair.

After the preference graph for the analyzed questions set is constructed, it is analyzed. The analysis is associated with the choice of the question to which all arcs lead to the top. This can be done in the absence of cyclic paths in the preference graph. In the presence of such, a root question close to the best variant is selected. The choice of the root question, therefore, becomes similar to how it is done for binary questionnaires [6, 7].

After the initially identified events $X = \{x_1, x_2, \dots, x_m\}$ set is divided into three subsets, proceed to solve the task of separating the events of each of the subsets. To do this, from the questions set $Y = \{y_1, y_2, \dots, y_n\}$ choose those questions that make sense for the corresponding events subset. Then preference graphs are constructed for each such questions subset and the root question is selected again.

The procedures are repeated until all identifiable events are separated into the questionnaire.

D. Optimization algorithm

The optimization algorithm for ternary questionnaires, derived from the root question method, contains the following items:

1. Pairwise comparison of questions is carried out.
2. Determines whether it is possible to establish relationships between all questions? If not, then another optimization method is selected, if yes – the transition to the next item of the algorithm is performed.
3. Pairs of questions are formed.
4. For each pair, a preference function $\Phi(A, B)$ is determined and the most preferred question is set.
5. Construction a preference graph.
6. The preference graph is analyzed, and the root question is selected.
7. The root question is set, and the initial set of events is divided into subsets.
8. The resulting subsets of the identified events are analyzed, and questions are identified that have meaning for each of them.
9. Are all identifiable events separated? If not, then steps 3–8 are repeated for each of the subsets of unshared events. If yes, then the required questionnaire is built.
10. End of the algorithm.

The main steps of this algorithm are related to defining comparison relations for each questions pair, calculating the preference function $\Phi(A, B)$ value for each question pair A and B , building the preference graph and choosing the root question. These procedures are performed sequentially from the separation of the full events set to the separation of each of the resulting subsets.

V. AN EXAMPLE OF OPTIMIZING THE TERNARY QUESTIONNAIRE

Many questions $Y = \{y_1, y_2, y_3, y_4\}$ are asked, allowing us to separate nine events of $X = \{x_1, x_2, \dots, x_9\}$. Baseline data, including the identified events weighting coefficients and the questions implementation costs, are given in matrix form (Tabl. 2). It is required to solve the task of constructing an optimal questionnaire.

TABLE II. QUESTIONNAIRE CHECK-LIST FOR THE ORIGINAL QUESTIONNAIRE

y_i	$c(y_i)$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
y_1	2	0	0	0	0	1	2	2	2	2
y_2	3	0	0	0	1	1	1	2	2	2
y_3	4	0	0	1	1	1	1	1	2	2
y_4	5	0	1	1	1	1	1	1	1	2
$p(x_i)$		0.01	0.01	0.05	0.2	0.4	0.3	0.01	0.01	0.01

Following the optimization algorithm, we construct a preference graph (Fig. 15).

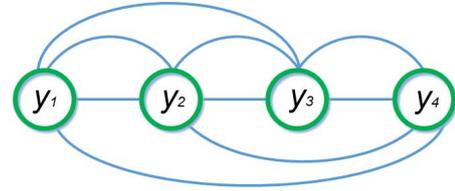


Fig. 15. Preference graph for the example in question.

Calculate the inequality for each pair of comparable issues Tabl. 2. Determine the value of preference functions for questions y_1 and y_2 :

$$\begin{aligned} \Phi(y_1, y_2) &= \frac{c(y_1) + c(y_2) \sum_{p_j \in X(y_{1m})} p_j}{c(y_2) + c(y_1) \sum_{p_i \in X(y_{2k})} p_i} = \\ &= \frac{2 + 3 \cdot 0.1}{3 + 2 \cdot 0.4} = 0.61. \end{aligned}$$

Since the value of $\Phi(y_1, y_2) < 1$, the question y_1 is preferable to the question y_2 : $y_1 < y_2$.

Similarly, we conclude the following: $y_2 < y_3$, $y_1 < y_4$, $y_1 < y_3$, $y_2 < y_4$, $y_3 < y_4$.

All questions that are in a comparison relationship can be summarized into one preference graph (Fig. 16). The arcs in the graph indicate a pairwise comparison of the questions; moreover, the arc is at the top of the question y_i , which is preferable to the question y_j . From the preference graph, the root question should be the question y_1 .

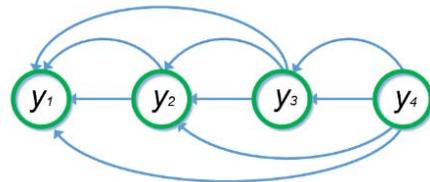


Fig. 16. Preference graph.

Perform the first partition of the events original set (Fig. 17).

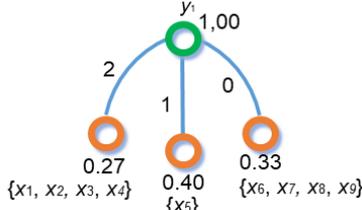


Fig. 17. The root question for the questionnaire under construction.

Now consider the resulting subsets $\{x_1, x_2, x_3, x_4\}$ and $\{x_6, x_7, x_8, x_9\}$. Define the root question for each of them.

All three remaining questions make sense with respect to the first and second subsets. Calculations show that the choice of the same question y_2 is most preferable. The questionnaire takes the form shown in Fig. 18.

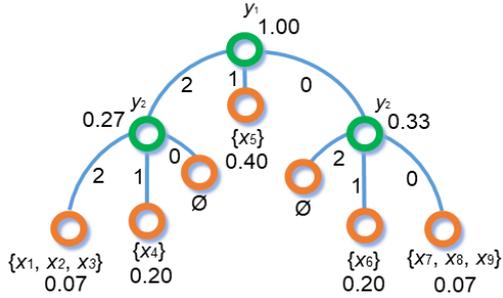


Fig. 18. Sub-questionnaire with the first two ranks questions.

Considering the subsets $\{x_1, x_2, x_3\}$ and $\{x_7, x_8, x_9\}$, we note that they are split by questions y_3 or y_4 . In both cases, the question of y_3 is preferable. The resulting two-element subsets in the next stage are split by the only remaining question y_4 . The result of the optimization is shown in Fig. 19.

The optimal questionnaire implementation cost is:

$$C = \sum_{i=1}^n p(y_i)c(y_i) = 2 \cdot 1.00 + 3 \cdot 0.27 + 3 \cdot 0.33 + 4 \cdot 0.07 + 4 \cdot 0.03 + 5 \cdot 0.02 + 5 \cdot 0.02 = 4.40.$$

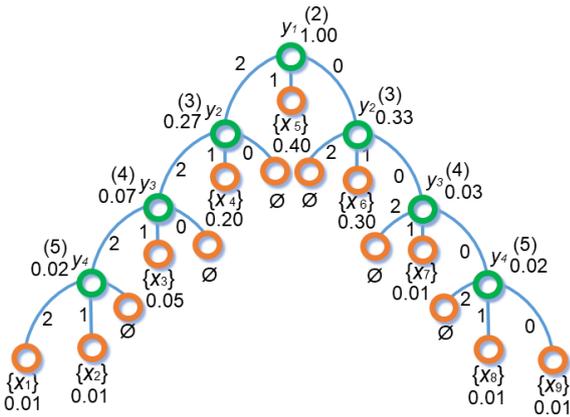


Fig. 19. Optimal questionnaire.

The given optimal questionnaire is obtained by using the branches and boundaries method and the dynamic programming method.

VI. TERNARY QUESTIONNAIRES WITH FAULTS

Polynomial Consider a special type of ternary questionnaire, the responses of which may contain faults. In this case, the initial set of events $X = \{x_1, x_2, \dots, x_m\}$ is divided into subsets of $X_{y_i}^2, X_{y_i}^1$ и $X_{y_i}^0$:

$$X_{y_i}^2 \cap X_{y_i}^1 = X_{y_i}^{12}, X_{y_i}^2 \cap X_{y_i}^0 = X_{y_i}^{20}, X_{y_i}^1 \cap X_{y_i}^0 = X_{y_i}^{10}. \quad (19)$$

At least one of the intersections subsets of the outcome's subsets $X_{y_i}^2, X_{y_i}^1$ or $X_{y_i}^0$ should not be empty, which is expressed by the condition:

$$W_{y_i} = X_{y_i}^{12} \cup X_{y_i}^{20} \cup X_{y_i}^{10} \neq \emptyset. \quad (20)$$

Formula (20) implies that for any event $x_j \in W_{y_i}$, the question outcome y_i is ambiguous. Depending on which of the subsets $X_{y_i}^{12}, X_{y_i}^{20}$ or $X_{y_i}^{10}$ are non-empty, it is possible to determine between which outcomes ambiguity arose. If there is no reservation as to between which outcomes of the question errors are possible, and we are talking about faults in general, then in order to designate the events included in the intersections of the $X_{y_i}^2, X_{y_i}^1$ or $X_{y_i}^0$ we can enter the designation «×». Thus, the questionnaire check-list element $b_{ij} = \times$ than $x_j \in W_{y_i}$.

An example of a questionnaire check-list, including questions with faults, is given in Tabl. 3.

TABLE III. TERNARY QUESTIONNAIRES CHECK-LIST WITH FAULTS

y_i	$c(y_i)$	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y_1	$c(y_1)$	2	2	×	1	0	0	0
y_2	$c(y_2)$	2	1	0	×	×	1	1
y_3	$c(y_3)$	×	1	0	2	1	0	×
y_4	$c(y_4)$	2	×	0	0	1	2	0
y_5	$c(y_5)$	×	×	0	2	1	0	2
y_6	$c(y_6)$	0	1	1	0	2	2	2
$p(x_j)$		$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_4)$	$p(x_5)$	$p(x_6)$	$p(x_7)$

A questionnaire is logically complete if it can be constructed with a questionnaire that distinguishes all events provided that any pair of (x_{j_1}, x_{j_2}) events is divided into different subsets with at least one question y_i . This condition can be written as follows:

$$\forall x_{j_1}, x_{j_2} \in X |_{j_1 \neq j_2}: \exists y_i: (x_{j_1} \in X_{y_i}^k) \& (x_{j_2} \notin X_{y_i}^k). \quad (21)$$

To assess the logical completeness of the questionnaire check-list use the difference matrix [4].

The difference matrix is the n by C_m^2 , Boolean matrix $\|d_{kj_1j_2}\|$ which elements $d_{kj_1j_2} = 1$ in case if question y_i distinguishes a couple of events (x_{j_1}, x_{j_2}) , that means:

$$d_{kj_1j_2} = 1 \Leftrightarrow (x_{j_1} \in X_{y_i}^k) \& (x_{j_2} \notin X_{y_i}^k). \quad (22)$$

The index k is the question number, the indices j_1 and j_2 are pairs of shared events.

If all matrix columns $\|d_{kj_1j_2}\|$ contain at least one a 1, then the questionnaire check-list is logically complete. If a pair of events (x_{j_1}, x_{j_2}) is not distinguished by the question, then the column will remain with zeros.

To build a difference matrix in the absence of faults in the answers, a ternary logic function would be appropriate, given by the truth table (Tabl. 3). This function $f = 1$ if the ternary variables $a \neq b$. This function performs a comparison operation: $f = a\Delta_{\neq}b$.

The function described above needs to be added due to the presence of questions with faults since the fourth type appears for the outcomes – the wrong answer «×». Define the comparison operation so that the condition is met:

$$f = a\Delta_{\neq} \times = 0. \quad (23)$$

TABLE IV. BOOLEAN ALGEBRA FUNCTION FOR CREATING A DIFFERENCE MATRIX

a	b		
	2	1	0
2	0	1	1
1	1	0	1
0	1	1	0

An analysis of a given questionnaire (Tabl. 3) showed that it is logically complete.

Using the comparison function allows you to build a difference matrix and establish the questionnaire completeness. Using the difference matrix, one can estimate the questionnaire redundancy. If it is possible to remove any question from it while preserving the completeness, then the questionnaire is redundant. The description of procedures for evaluating redundancy is like the case of binary questionnaires analyzing redundancy described in [4, 7].

Consider the procedure for design a questionnaire using the questionnaire check-list in Fig. 20. In this questionnaire, there are questions that have the wrong answer. For example, let question y_2 be selected as the root question. Posing this question on a set of events breaks it into three subsets. An event for which the possibility of faults is implied is included in the subsets of all three outcomes. The subsets are formed $X_{y_2}^2 = \{x_1, x_4, x_5\}$, $X_{y_2}^1 = \{x_2, x_4, x_5, x_6, x_7\}$ and $X_{y_2}^0 = \{x_3, x_4, x_5\}$. The set $W_{y_2} = \{x_4, x_5\}$ is included in each of the root question outcomes. The subsets $X_{y_2}^2$ and $X_{y_2}^0$ separated by questions y_4 and y_3 . The subset $X_{y_2}^1$ separated by questions y_5 . These examples are not the only ones that can be seen by analyzing a given questionnaire check-list. The question y_5 on the obtained subset $X_{y_2}^1$ also, have a wrong answer when identifying the event x_2 . It is included in each of the outcome's subsets of this question. The partitioning procedure is repeated until all the events that can be divided are separated.

The received questionnaire main feature is the identifying possibility of the same event along with several routes in a column at the same time. Also, the difference is the number of the possible questionnaire which significantly increases due to the presence of different identification options for the same event.

Changing the questionnaire regarding its usual form, when each event is identified by a separate route, leads to the need to change the method of determining the average implementation costs.

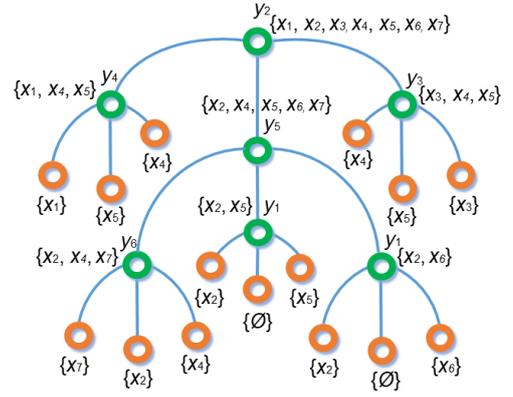


Fig. 20. Ternary questionnaires with faults.

For each event $x_j \in W_{y_i}$, as in [7], we introduce an additional weighting function $\delta_{j,i}^k$, $k = 2,1,0$, such that:

$$\delta_{j,i}^k \in [0; 1]; \quad \sum_k \delta_{j,i}^k = 1. \quad (24)$$

The function $\delta_{j,i}^k$ can be assigned a probabilistic meaning: the value of $\delta_{j,i}^k$ is the probability that in the state x_j to the question y_i will be the answer «k». If in practical applications situations arise when the answer to a question no probabilistic content has, then we will assume that the outcome of such questions is fuzzy sets [15]. In such situations, the determinateness degree of the question outcome y_i is determined by the values:

$$\lambda_{j,i}^k = \frac{|W_{y_i}|}{|X_{y_i}^k|}, \quad k = 2,1,0. \quad (25)$$

The functions values $\delta_{j,i}^k$ will be the belonging measures values of the event x_j to k to the end of the question y_i .

Introducing additional weight functions for a subset of events $x_j \in W_{y_i}$, replace the expression to calculate the cost of identifying an event x_j :

$$C(x_j) = \sum q_{\tau}(x_i)c_{\tau}(x_i), \quad (26)$$

where $c_{\tau}(x_i)$ – identification cost of the set x_j on the route τ in the questionnaire, a $q_{\tau}(x_i)$ – event identification probability x_j along the route τ in the questionnaire.

The values $q_{\tau}(x_i)$ in (25) determined by the formula:

$$q_{\tau}(x_i) = \prod_{y_i} \delta_{j,i}^k. \quad (27)$$

In (27) multiply all quantities $\delta_{j,i}^k$ for all questions, owned route τ in the questionnaire, and the outcomes subset index k is determined based on the answer to question y_i . If the answer to the question is deterministic, then $\delta_{j,i}^k = 1$. The questionnaire cost, in this case, can be determined by the formula:

$$C = \sum_{j=1}^m p(x_j)C(x_j). \quad (28)$$

The formula (28) is an analogue of the formula (3). The difference between formulas (3) is that it calculates the event identification cost by routes and formula (28) uses the average event identification cost obtained by formula (27).

Optimization of questionnaires with questions that have faults, is as follows.

The original questionnaire, including questions with faults, is expanded by introducing three copies of events $x_j^{k,i}$, $k = 2,1,0$ for each of the questions y_i .

If there are several questions for which event x_j in the questionnaire is indicated by « \times », then three copies of $x_j^{k,i}$ are entered for each of the questions.

If in the x_j state the answer to the question y_i was outcome « k », then in x_j^k states the answers will also be outcomes « k ».

For each new event, normed weighted functions are introduced:

$$p(x_j^{k,i}) = p(x_j) \prod_{k=1}^K \delta_{j,i}^k, \quad k = 2,1,0. \quad (29)$$

where $\delta_{j,i}^k$ – the probability of answer k to question y_i in state x_j ; K – the number of questions that were not answered correctly.

If the questionnaire is optimized using only the above steps, the questionnaire will be optimal and will share all entered events $x_j^{k,i}$, $k = 2,1,0$, it will not be optimal for the initial questionnaire with questions that make faults. In this case, the received questionnaire may also include questions that share a pair of newly introduced events, copies of some event x_j . Such questions will not make sense. To obtain an optimal solution for the initial questionnaire, it is necessary to expand the set of questions $Y = \{y_1, y_2, \dots, y_n\}$, by entering into it a set of imaginary questions Y^ϕ , whose number is determined by the number of events in the set $X = \{x_1, x_2, \dots, x_m\}$, for which there is at least one question that makes faults.

Such questions will not be ternary but will have a base $3^K + 1$. The value of the element b_{ij} in the expandable questionnaire check-list is determined by the number i in the event $x_j^{k,i}$.

For imaginary questions from the set Y^ϕ the cost is determined to be zero and we agree that the question $y_i \in Y^\phi$ can be asked only if the zero outcomes of the imaginary question does not contain a single event:

$$X_{y_i}^0 = \emptyset, y_i \in Y^\phi. \quad (30)$$

Further, the optimization of the received questionnaire with fault correction by the known methods [4, 7], with the only difference that at each stage, for imaginary questions, it is necessary to check the condition (30). After the optimization procedure is completed, all imaginary questions $y_i \in Y^\phi$ are discarded.

Previously, the authors considered questionnaires with questions allowing for undefined faults. Under these conditions, it is not known which of the three outcomes $X_{y_i}^2$, $X_{y_i}^1$ or $X_{y_i}^0$ of any question y_i can belong to an event. In reality, for a number of questions, there may be a limitation in between what outcomes of each question a fault is possible. This simplifies the procedure for determining the values in the questionnaire check-list and its adjustment in comparison with the method of interpreting any faults as a transition to an arbitrary subset of the question outcomes.

VII. TERNARY QUESTIONNAIRES WITH DON'T CARES

In a few separating and identifying event tasks, there are situations in which the value of a condition is not essential

for fulfilling a rule. Examples of such tasks are binary decision check-list [7]. If the physical formulation of the tasks, the mathematical model of which is the questionnaire, allows an arbitrary deterministic answer, then it is considered undefined and is indicated by “–”. Consider ternary questionnaires with undefined answers. As an example, consider the questionnaire check-list given in Tabl. 5.

TABLE V. TERNARY QUESTIONNAIRES CHECK-LIST WITH AN INDEFINITE ANSWER

y_i	$c(y_i)$	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y_1	$c(y_1)$	2	2	–	1	0	0	0
y_2	$c(y_2)$	2	1	0	–	–	1	1
y_3	$c(y_3)$	–	1	0	2	1	0	–
y_4	$c(y_4)$	2	–	0	0	1	2	0
y_5	$c(y_5)$	–	–	0	2	1	0	2
y_6	$c(y_6)$	0	1	1	0	2	2	2
	$p(x_j)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_4)$	$p(x_5)$	$p(x_6)$	$p(x_7)$

The questionnaire will be logically complete if for any pair of events (x_{j_1}, x_{j_2}) there exists at least one question y_i with outcomes determined for this pair, such that both events belong to its different outcomes. A questionnaire with undefined answers may turn out to be incomplete but allow for the completeness of the outcomes of undefined answers to be logical completeness.

In order to design a questionnaire on a questionnaire check-list with uncertainties, it is necessary to arbitrarily add them if the questionnaire was logically complete, and to deliberately add them, considering the completeness of the converted questionnaire. If we are talking about building a questionnaire with a minimum implementation cost, the task becomes more complicated: a “special” definition of the questionnaire is required.

A questionnaire can be determined by the number of ways 3^γ , where γ – the number of don't cares. By designing all possible questionnaires and choosing the one whose implementation cost is the lowest, we will arrive at the desired result. However, the number of search options is significant. Even for the example under consideration in table 6, it is $3^8 = 6561$. In [7] it is noted that there is no efficient algorithm with polynomial complexity for solving this task therefore, the tasks should be solved by approximate methods. Consider one of these methods.

It is necessary to define the questionnaire check-list rows, which bear the “–”. For these purposes, a difference matrix is constructed. At the same time, when the “–” appears in the comparison, a “–” s is placed in the difference matrix. After that, the difference matrix is analyzed and there are such columns where there are no single values. The values at the place of the “–” are determined in such a way that there exists at least one question y_i separating the corresponding pair of events (x_{j_1}, x_{j_2}) in the difference matrix column. If for one question there are several options for complementing (several pairs of events), then when determining the questionnaire check-list, they try to divide a couple of events with the maximum total probability $p_{j_1, j_2} = p_{j_1} + p_{j_2}$. This approach in [7] for binary questionnaires. It also says that this method allows you to

get solutions with a quick determination of answers to questions. Also, to ensure the design of a questionnaire with an implementation cost close to the minimum.

VIII. CONCLUSION

Ternary questionnaires that are of the type polychotomous, are a more general case than binary questionnaires. Although any questionnaire can be relatively easily converted to a binary questionnaire, in some tasks this is not required. Moreover, the binary questionnaire will contain a larger number of questions than the corresponding ternary questionnaire. Depending on the ternary and binary questionnaires questions implementation cost, this factor will influence the final questions implementation cost. If its value is set to a limit, in some cases the use of polychotomic questionnaires may be justified.

Ternary questionnaires are the next class of homogeneous questionnaires in terms of the basis of the questions. Their use and research give the opportunity to further study the use of other homogeneous (and considering fictitious outcomes – and heterogeneous) questionnaires.

The proposed optimization algorithm based on the root question method makes it possible to obtain the best ternary questionnaires at the implementation costs. An optimization algorithm based on this method has a polynomial complexity estimate [7]. However, this method is not always applicable, but only if it is possible to establish comparison relations between all the questions in the questionnaire. If a comparison relationship cannot be established, the classical branch and bound methods and dynamic programming should be used to optimize ternary questionnaires.

The questionnaires application practical side examined in this paper is associated with many applications, one of which is technical diagnostics. Despite the widespread use of binary questionnaires for the construction of diagnostic algorithms in several tasks it is necessary to use other types of questionnaires (this is determined by the valid questions) [11]. Despite the ease of converting any questionnaires into binary questionnaires, the latter have a significant drawback - using them to identify the same set of events will require more questions than using other types of questionnaires (including ternary). Depending on the question's implementation cost of the ternary and binary questionnaires, this factor will influence the final implementation cost. If its value is set to a limit, in some cases the use of homogeneous (and, by the way, heterogeneous) questionnaires may be justified.

IX. ACKNOWLEDGEMENTS

The paper authors express thanks to Pavel P. Parkhomenko, DSc, corresponding member of the Russian Academy of Sciences, Senior Research Associate of Labora-

tory of Technical Diagnostics and Fault Tolerance of V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, as well as Galina P. Aksenova, senior researcher at the same scientific institution, for discussing the results and general positive feedback on the work done.

REFERENCES

- [1] R. Ubar, J. Raik, and H.-T. Vierhaus "Design and Test Technology for Dependable Systems-on-Chip (Premier Reference Source)", Information Science Reference, Hershey – New York, IGI Global, 2011, 578 p.
- [2] V. Hahanov "Cyber-Physical Computing for IoT-driven Services", New York, Springer International Publishing AG, 2018, 279 p., doi: 10.1007/978-3-319-54825-8.
- [3] P.P. Parkhomenko "Theory of Questionnaires (Review)", Automation and Remote Control, 1970, vol. 31, Issue 4, pp. 639-655.
- [4] P.P. Parkhomenko, and E.S. Sogomonyan "Technical Diagnosis Fundamentals (Diagnostic Algorithm Optimization, Apparatus Means)" (in Russ.), Moscow: Energoatomizdat, 1981, 320 p.
- [5] C.F. Picard "Graphs and Questionnaires", Netherlands: North-Holland Publishing Company, 1980, 431 p.
- [6] A.Yu. Arzhenenko, O.G. Kazakova, and B.N. Chugaev "Optimization of Binary Questionnaires", Automation and Remote Control, 1985, Vol. 46, issue 11, pp. 1466-1472.
- [7] A.Yu. Arzhenenko, and B.N. Chugaev "Optimal Binary Questionnaires" (in Russ.), Moscow: Energoatomizdat, 1989, 128 p.
- [8] A.Yu. Arzhenenko, and V.A. Vestyak "Modifying the Tolerant Substitution Method in Almost Uniform Compact Surveys", Automation and Remote Control, 2012, Vol. 73, Issue 7, pp. 1195-1201, doi: 10.1134/S0005117912070090.
- [9] B.N. Chugaev, and A.Yu. Arzhenenko "Optimal Identification of Random Events" (in Russ.), Economics, Statistics and Informatics, 2013, Issue 2, pp. 188-190.
- [10] P.P. Parkhomenko "Questionnaires and Organizational Hierarchies", Automation and Remote Control, 2010, Vol. 71, issue 6, pp. 1124-1134, doi: 10.1134/S0005117910060135.
- [11] D.V. Efanov, V.V. Khoroshev, G.V. Osadchy, and A.A. Belyi "Optimization of Conditional Diagnostics Algorithms for Railway Electric Switch Mechanism Using the Theory of Questionnaires with Failure Statistics", Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS'2018), Kazan, Russia, September 14-17, 2018, pp. 237-245, doi: 10.1109/EWDTS.2018.8524620.
- [12] G. Duncan "Heterogeneous Questionnaire Theory", SIAM Journal on Applied Mathematics, 1974, Vol. 27, Issue 1, pp. 59-71, doi: 10.1137/0127005.
- [13] A.H. Land, and A.G. Doig "An Automatic Method of Solving Discrete Programming Problems", Econometrica, 1960, Vol. 28, No. 3, pp. 497-520.
- [14] R.E. Bellman "Dynamic Programming", Princeton University Press, Princeton NJ, 1957, 392 p.
- [15] L.A. Zadeh "Probability Measures of Fuzzy Events", Journal of Mathematical Analysis and Applications, 1968, Vol. 23, Issue 2, pp. 421-427, doi: 10.1016/0022-247X(68)90078-4.

Harmonic Distortions in Analog Interfaces Based on Differential Difference Amplifiers

Nikolay V. Butyrlagin
Information system and
radioengineering
Don State Technical University
Rostov-on-Don, Russia
butyrlagin@gmail.com

Anna V. Bugakova
Information system and
radioengineering
Don State Technical University
Rostov-on-Don, Russia
annabugakova.1992@mail.ru

Nikolay N. Prokopenko
Information system and
radioengineering
Don State Technical University,
IPPM RAS
Rostov-on-Don, Russia
prokopenko@sssu.ru

Mikhail F. Mitsik
Mathematics and applied informatics
Don State Technical University
Rostov-on-Don, Russia
m_mits@mail.ru

Alexey E. Titov
Automatic Control Systems
South Federal University
Rostov-on-Don, Russia
alex.evgeny.titov@gmail.com

Abstract— The article considers a mathematical analysis and computer simulation of total harmonic distortion (THDf) in CMOS differential difference amplifiers (DDA). We have presented THDf's dependence and sensitivity of output voltage in typical DDA circuit to input differential stages (DS) limiting voltage and their static current mode. We have found, that DDA typical circuit's THDf is significantly worse at microamp currents of input CMOS transistors.

Keywords—differential difference amplifier, input differential stage, total harmonic distortion, transfer characteristic, limiting voltage, CMOS transistors

I. INTRODUCTION

DDA [1-2] are among up-to-date active analog components, numbered more than 200 modifications [3-6]. They have some advantages in comparison with classical operational amplifiers (OPAMP) [7-10]. But this subclass of analog component base is not analyzed enough in the context of total harmonic distortion's (THDf) estimation at low current consumption [11].

The purpose and novelty of the paper are mathematical analysis and computer simulation of THDf in CMOS DDA, including situations with different static current modes of its input differential stages (DS1, DS2).

II. TERMS OF REFERENCES

There is a circuit of typical DDA with two input non-identical (generally) differential stages (DS1, DS2) on Fig. 1

The circuit includes DS1, DS2, buffer amplifier (BA), common negative feedback resistors R1, R2 and balancing capacitor C_b. The sensors' output signals are fed to the differential input of DS1 (In.1, In.2).

The output currents of CMOS DS1 and DS2 are approximated closely (Fig. 1) by hyperbolic tangent and defined by the following formulas:

$$i_{out.1} = I_{01} \operatorname{th} \frac{v_{12}}{V_{lim.1}}, \quad i_{out.2} = I_{02} \operatorname{th} \frac{v_{34}}{V_{lim.2}}, \quad (1)$$

where $V_{lim.1}$, $V_{lim.2}$, I_{01} , I_{02} are parameters of DS1, DS2 transfer characteristics (TC) (Fig. 2), when its linear piecewise approximation, v_{12} is DS1 differential stage's output voltage. In particular case the voltage is the following, when variable signals at frequency f_c : $v_{12} = V_{12} \sin 2\pi f_c$.

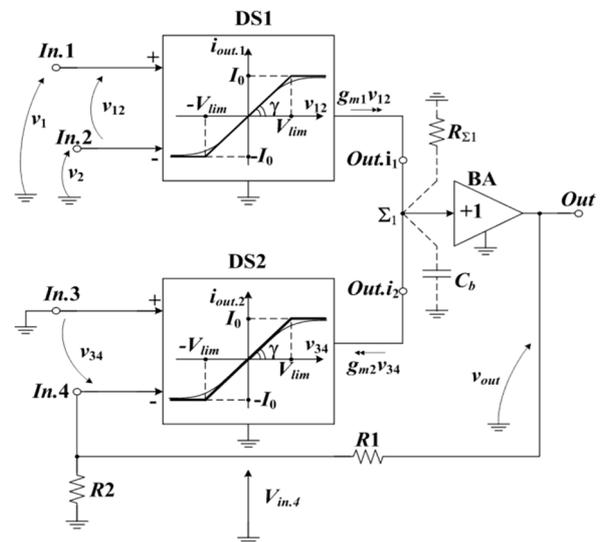


Fig. 1. DDA's functional circuit with non-linear input stages DS1, DS2.

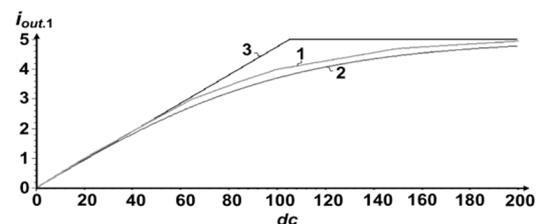


Fig. 2. Comparison of transfer characteristics (TC) for XFab DS (1) with its hyperbolic (2) and linear piecewise (3) approximation.

There are numerical values of maximum relative error δ_{max} of TC's approximation for two CMOS processes, that are SG25VD and XFab XB06, on Table 1.

TABLE I. CMOS DS'S TRANSFER CHARACTERISTIC'S APPROXIMATION ERROR WITH FUNCTION TH X (1)

	Limiting Voltage $V_{lim.1}$, mV	Maximum Relative Error $\delta_{max} = \Delta I/I_{01}$	Static Current I_{01} (DS1, DS2), μA
Process SiGe25VD			
1	85	5%	1
2	105	6%	10
3	158	9%	100
4	600	14%	1000
Process XFab06			
5	105	6%	1
6	130	7%	10
7	230	8%	100
8	600	12%	500
9	1000	14%	1000

It can be shown, that DDA's output voltage (Fig. 1) V_{out} at hyperbolic approximation is related to input voltage v_{12} by the following non-linear expression:

$$v_{out} = \left(1 + \frac{R_1}{R_2}\right) \cdot \frac{V_{lim.2}}{2} \cdot \ln \left(\frac{I_{02} + I_{01} \operatorname{th} \left(\frac{v_{12}}{V_{lim.1}} \right)}{I_{02} - I_{01} \operatorname{th} \left(\frac{v_{12}}{V_{lim.1}} \right)} \right). \quad (2)$$

Formula (2) allows connecting THdf and DS1 and DS2 transfer characteristic parameters ($V_{lim.1}$, $V_{lim.2}$, I_{01} , I_{02}), in particular considering their variety.

III. DDA BASE CIRCUIT'S BASE PARAMETERS' SENSITIVITY FUNCTION

There is a practical interest to define DDA's (Fig. 1) sensitivity to change of DS1, DS2 input differential stages' TC's base parameters. According to the classical definition of sensitivity [12] it is:

$$S_x^{F(x)} = \frac{dF(x)}{dx} \cdot \frac{x}{F(x)},$$

By rearrangement of equation (2) one deduces, that the sensitivity V_{out} to parameter I_{01} is the following:

$$S_{I_{01}}^{V_{out}} = \frac{\operatorname{th} \left(\frac{v_{12}}{V_{lim.1}} \right)}{\left(\frac{I_{02} - I_{01}}{I_{01} - I_{02}} \cdot \operatorname{th}^2 \left(\frac{v_{12}}{V_{lim.1}} \right) \right) \cdot \left(\operatorname{arth} \left(\frac{I_{01}}{I_{02}} \cdot \operatorname{th} \left(\frac{v_{12}}{V_{lim.1}} \right) \right) \right)}$$

Similarly, the current I_{02} sensitivity V_{out} is :

$$S_{I_{02}}^{V_{out}} = \frac{\operatorname{th} \left(\frac{v_{12}}{V_{lim.1}} \right)}{\left(\frac{I_{02} - I_{01}}{I_{01} - I_{02}} \cdot \operatorname{th}^2 \left(\frac{v_{12}}{V_{lim.1}} \right) \right) \cdot \left(\operatorname{arth} \left(\frac{I_{01}}{I_{02}} \cdot \operatorname{th} \left(\frac{v_{12}}{V_{lim.1}} \right) \right) \right)}$$

The sensitivity V_{out} to limiting voltage $V_{lim.1}$ of first input stage DS1 is:

$$S_{V_{lim.1}}^{V_{out}} = - \frac{v_{12} \cdot \operatorname{sech}^2 \left(\frac{v_{12}}{V_{lim.1}} \right)}{V_{lim.1} \cdot \left(\frac{I_{02} - I_{01}}{I_{01} - I_{02}} \cdot \operatorname{th}^2 \left(\frac{v_{12}}{V_{lim.1}} \right) \right) \cdot \left(\operatorname{arth} \left(\frac{I_{01}}{I_{02}} \cdot \operatorname{th} \left(\frac{v_{12}}{V_{lim.1}} \right) \right) \right)}$$

The sensitivity V_{out} to limiting voltage $V_{lim.2}$ of DS2 second input stage DS2 is: $S_{V_{lim.2}}^{V_{out}} = 1$.

The numerical computer simulation of the above sensitivity functions (Fig. 3, Fig. 4) has shown, that DDA with architecture of Fig. 1 at low V_{lim} is characterized by increased influence of DS1, DS2 ($V_{lim.1}$, $V_{lim.2}$, I_{01} , I_{02}) transfer characteristics' parameters' nonidentity to dependence $V_{out} = f(V_{in})$.

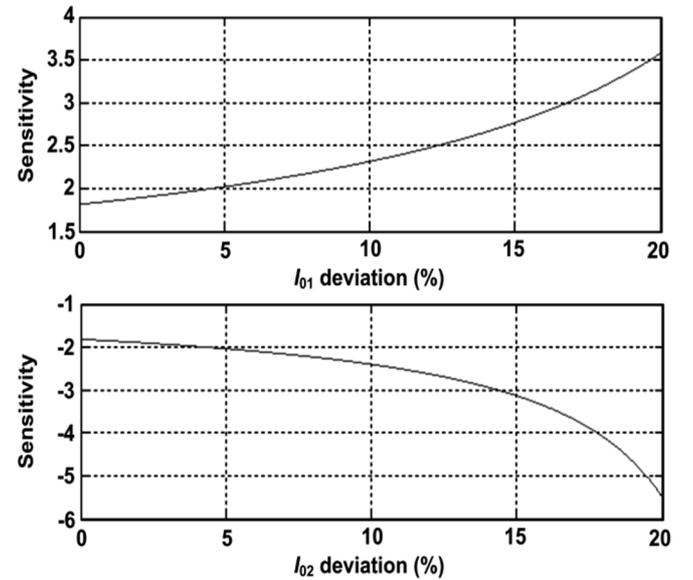


Fig. 3. Sensitivities V_{out} to I_{01} and I_{02} current change.

One has obtained, the curves (Fig. 3) of sensitivity functions $S_{I_{01}}^{V_{out}}$ and $S_{I_{02}}^{V_{out}}$, when $V_{lim.1} = V_{lim.2} = 1$ V, $V_{12} = 1$ V and current deviations $I_{01} = I_{02} = 300$ μA relative to each other are up to 20% correspondently. So even a little change of DS1 input stage current I_{01} relative to other DS2 input stage current I_{02} (and conversely) leads to significant increase of sensitivity V_{out} to changeable I_{0i} , where i is input stage number.

There is a dependence of sensitivity function $S_{V_{lim.1}}^{V_{out}}$ to voltage change $V_{lim.1}$ (0.5 V...2 V) at $V_{12} = 1$ V, $V_{lim.2} = 1$ V, $I_{01} = 300$ μA , $I_{02} = 270$ μA on Fig. 4. The graph of Fig. 4 shows, that a parameter $V_{lim.1}$ decrease, that is lower than input signal amplitude value V_{12} , leads to significant increase of sensitivity V_{out} to $V_{lim.1}$.

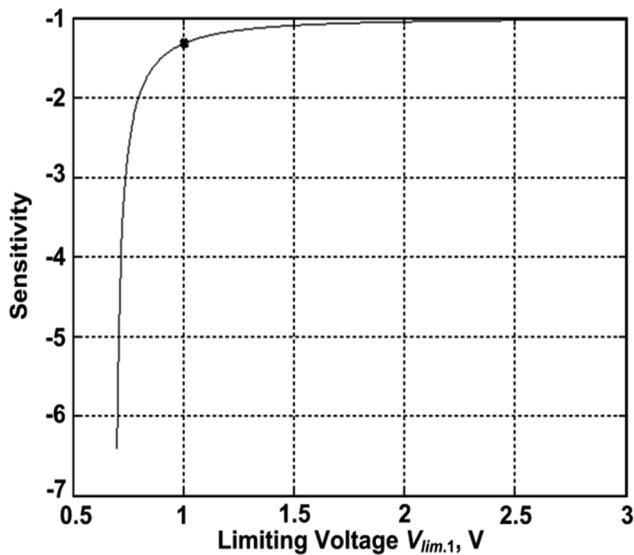


Fig. 4. Sensitivity V_{out} to change of parameter $V_{lim,1}$.

IV. DDA TOTAL HARMONIC DISTORTION IN DDA BASE CIRCUIT

The equation (2) allows getting a dependence of THDF circuit (Fig. 1) total harmonic distortion from signal input amplitude V_{I2} at different relations I_{01}/I_{02} and limiting voltages $V_{lim,1}, V_{lim,2}$:

Situation 1. Suppose $V_{lim,1} = V_{lim,2} = 0.2 V$, $f_c = 1 \text{ kHz}$ (input signal frequency), $A_D = 1$ (closed DDA gain), $I_{01} = 300 \mu\text{A}$. There is a THDF dependence on amplitude V_{I2} at $I_{02} = 297 \mu\text{A}$, that is the deviation from I_{01} by 1% on Fig. 5. The dependence is marked by “+”. There is a THDF dependence on amplitude V_{I2} at $I_{02} = 270 \mu\text{A}$, that is the deviation from I_{01} by 10%. It is marked as “o”.

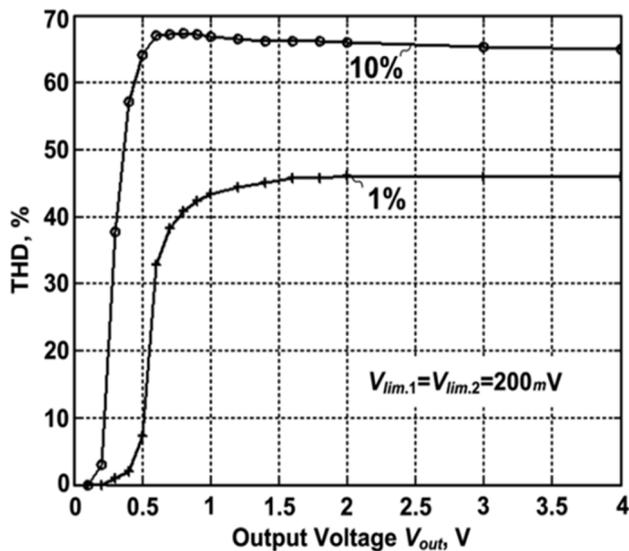


Fig. 5. THDF dependence on amplitude U_{I2} , situation 1.

Situation 2. Suppose, limiting voltages $V_{lim,1} = V_{lim,2} = 2 V$, $f_c = 1 \text{ kHz}$, $A_D = 1$ (DDA gain with

100% feedback), $I_{01} = 300 \mu\text{A}$. There is a THDF dependence on amplitude V_{I2} at $I_{02} = 297 \mu\text{A}$, that is the deviation from I_{01} by 1% on Fig. 6. The dependence is marked with “+”. There is a THDF dependence from amplitude V_{I2} at $I_{02} = 270 \mu\text{A}$, that is deviation from I_{01} by 10%. It is marked with “o”.

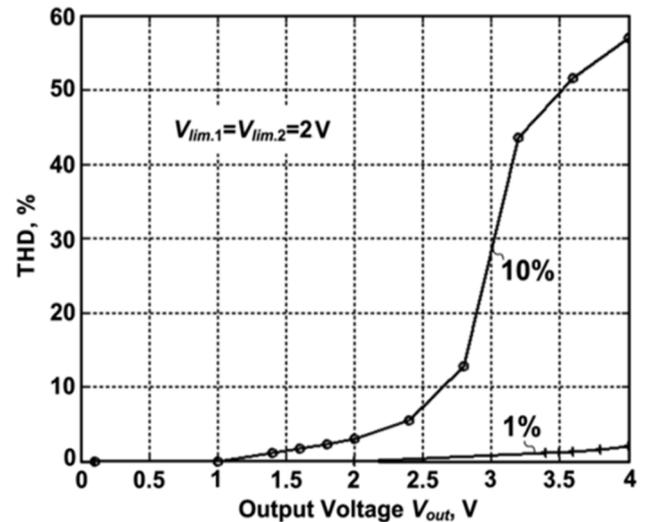


Fig. 6. THDF dependence on amplitude U_{I2} , situation 2.

So there is a significant decrease of THDF sensitivity in DDA circuit (Fig. 1) to current (I_{01}, I_{02}) nonidentity and limiting voltages $V_{lim,1}(V_{lim,2})$, when limiting voltages (of the order of 2 V) are higher.

V. DDA CMOS DS1, DS2 STATIC MODE'S INFLUENCE ON HARMONIC DISTORTION

The peculiarity of DDA on CMOS transistors is that their limiting voltages decreases to tens of millivolts at low static current [13-15]. The computer simulation of DDA CMOS THDF (Fig. 7) at low current and high current modes DS1 and DS2 (Fig. 8) has shown, that DDA output voltage spectrum (Fig. 7) is significantly enhanced in the circuit (Fig. 1) at micro-mode.

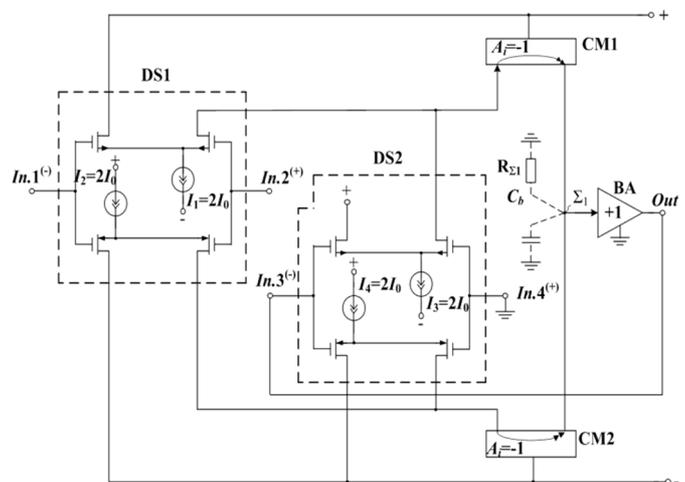
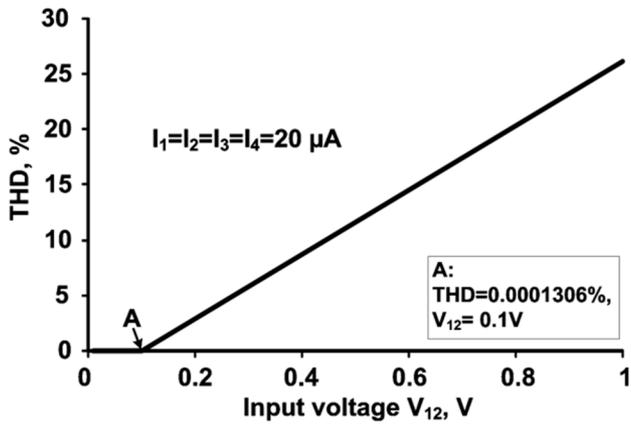
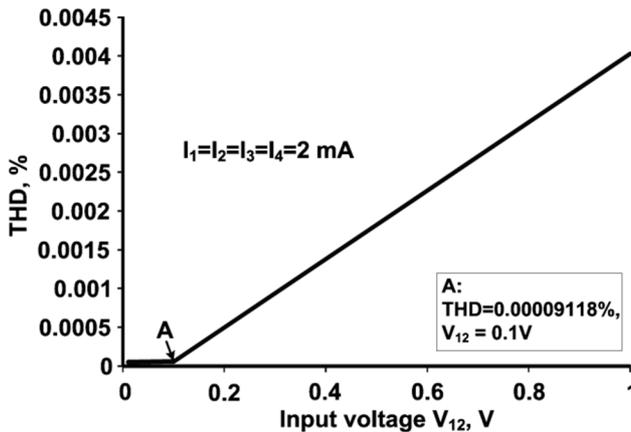


Fig. 7. CMOS XFab06 DDA with two input stages DS1, DS2.



a)



b)

Fig. 8. THDf dependence on input voltage amplitude V_{12} at different static currents of DS CMOS DDA (Fig. 7).

VI. RESULTS AND DISCUSSION

In order to provide small THDf during designing of analog interfaces with architecture of Fig. 1, one must select DS1, DS2 input stages circuitry, which should have high values of limiting voltages. Considering significant influence of DDA CMOS differential stages' static current, it is not recommended to apply micro-mode of CMOS field transistors in DS1, DS2 circuits.

VII. CONCLUSION

We have shown that there is a significant influence of static mode of current for input differential stages and their base parameters' variation I_{01} , I_{02} , $V_{lim.1}$, $V_{lim.2}$. on numerical values of DDA CMOS THDf in typical circuit.

The micropower DDA CMOS cannot have low values of THDf. To eliminate this contradiction, it is required to use local negative feedback resistors in differential stages (DS1, DS2), which extend their linear operating range. But in this case other DDA parameters (gain, common mode rejection ratio etc.) are significantly degraded. There is a need of special circuitry for DS1, DS2.

REFERENCES

- [1] "Design of Differential Amplifier Using Current Mirror Load in 90 nm CMOS Technology," Advances in Intelligent Systems and Computing, vol 862., Publisher: Springer., pp.421-429.
- [2] F. Khateb and T. Kulej, "Design and Implementation of a 0.3-V Differential Difference Amplifier," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 2, pp. 513-523, Feb. 2019. DOI: 10.1109/TCSI.2018.2866179.
- [3] H. H. Kuntman and A. Uygur, "New possibilities and trends in circuit design for analog signal processing," 2012 International Conference on Applied Electronics, Pilsen, 2012, pp. 1-9.
- [4] R. Senani, D. R. Bhaskar and A. K. Singh, "Current Conveyors: Variants, Applications and Hardware Implementations," Springer International Publishing, Switzerland, 2015. DOI 10.1007/978-3-319-08684-2.
- [5] D. Biolek, R. Senani, V. Biolkova, and Z. Kolka, "Active elements for analog signal processing: classification, review, and new proposals," Radioengineering, vol. 17, no. 4, pp. 15-32, 2008.
- [6] R. Senani, D. R. Bhaskar, V. K. Singh and R. K. Sharma, "Sinusoidal Oscillators and Waveform Generators using Modern Electronic Circuit Building Blocks," Springer International Publishing, Switzerland, January 2016. DOI: 10.1007/978-3-319-23712-1.
- [7] B. A. Minch, "A CMOS differential-difference amplifier with class-AB input stages featuring wide differential-mode input range," 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, 2017, pp. 1-4. DOI: 10.1109/ISCAS.2017.8050488.
- [8] R. S. AlDisi and F. T. Jaber, "Theoretical investigation on the effect of individual stage-gain selection on the 3-dB bandwidth of three-op-amp difference amplifiers," 2016 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Batu Feringghi, 2016, pp. 155-158. DOI: 10.1109/ISCAIE.2016.7575055.
- [9] J. S. Mincey, C. Briseno-Vidrios, J. Silva-Martinez and C. T. Rodenbeck, "Low-Power G_m -C Filter Employing Current-Reuse Differential Difference Amplifiers," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 64, no. 6, pp. 635-639, June 2017. DOI: 10.1109/TCSII.2016.2599027.
- [10] A. J. Lopez-Martin, M. P. Garde and J. Ramirez-Angulo, "Class AB differential difference amplifier for enhanced common-mode feedback," in Electronics Letters, vol. 53, no. 7, pp. 454-456, 30 3 2017. DOI: 10.1049/el.2017.0347.
- [11] Yongwang Ding ; R. Harjani, "A +18 dBm IIP3 LNA in 0.35 μm CMOS," 2001 IEEE International Solid-State Circuits Conference. Digest of Technical Papers. ISSCC, pp. 1-3. DOI: 10.1109/ISSCC.2001.912587.
- [12] A. V. Frolov "How calculate electric circuits' sensitivity to their components' parameters' change," TUSUR's Proceedings, 2012. No. 1 (25), Part I. pp. 29-33. (in Russian).
- [13] N.V. Butyrlagin, N.N. Prokopenko, E.M. Savchenko, A.S. Budyakov, "Design features of high-speed CMOS differential difference operational amplifiers at low static current consumption," 26th Telecommunications forum TELFOR 2018, Serbia, Belgrade, November 20-21, 2018. pp. 1-4. DOI: 10.1109/TELFOR.2018.8611965.
- [14] N.N. Prokopenko, N.V. Butyrlagin, A.V. Bugakova, "The Comparative Analysis of the Maximum Slew Rate of the Output Voltage BJT and CMOS (SiGe TSMC 0.35 μ) Operational Amplifiers," 19th International conference on micro/nanotechnologies and electron devices EDM 2018: Proceedings, Erlagol, Altai Republic, 29 June - 3 July, 2018, pp. 712-717. DOI: 10.1109/EDM.2018.8435058.
- [15] N.N. Prokopenko, A.V. Bugakova, P.S. Budyakov, A.I. Serebryakov, "Method for Speeding a Differential Operational Amplifier in the Invert Connection Circuit," Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2018), Kazan, Russia, September 14 - 17, 2018, pp.676-680.

Intelligent Sensor Measurement of GTE Gas Temperature with Thermistors

Zh. A. Sukhinets
Department of
telecommunication systems
Ufa State Aviation Technical
University
Ufa, Russia
sukhinets@mail.ru

A. I. Gulin
Department of automation of
technological processes and
production
Ufa State Petroleum
Technological University
Ufa, Russia
gulin1940@gmail.com

N. M. Safyannikov
Department of computer
engineering
Saint-Petersburg State
Electrotechnical University
“LETI”
Saint-Petersburg, Russia
NMSafyannikov@etu.ru

N. N. Prokopenko
Department Information systems
and radio engineering
Don State Technical University
Rostov-on-Don, Russia
prokopenko@sssu.ru

O. O. Valiamova
Department of automation of
technological processes and
production Ufa State Petroleum
Technological University
Ufa, Russia
oopez@mail.ru

O.I. Bureneva
Department of computer
engineering
Saint Petersburg Electrotechnical
University “LETI”
Saint Petersburg, Russia
oibur@mail.ru

Abstract— Based on the analysis of the existing schemes and measuring instruments of high temperatures GTE, the article proposes an intelligent system using the developed design of a thermistor made of tungsten-rhenium wire, allowing its installation on the jacks of previously used thermocouples. The calculations of the thermistor, taking into account the dependence of the resistivity change and temperature coefficient of resistance on temperature. Using the conversion functions, the optimal value of the output frequency of the measurement system for transmission over a two-wire communication line with high noise immunity is calculated.

Keywords— high temperature, thermistor, phasing chain, frequency, microcontroller

I. INTRODUCTION

Gas turbine engines (GTE) are widely used not only in aviation, where they are the main power plants of aircraft, but also in the gas industry [1] as gas compressor units. The gas temperature before the GTE turbine is one of the main parameters that determine the traction characteristics and engine life. Both direct and indirect methods are used to measure temperature. Direct measurement methods involve the use of temperature sensors, which thermoelements are usually used. Design problems of such sensors are connected [2] with the choice of heat-resistant materials and reducing the size. Indirect methods of measuring the temperature of the gas is also called the method of calculating the temperature. The calculation of temperature consists in measuring various parameters of the gas turbine engine, for example, the rotor speed and its derivative, and in determining the gas temperature on the basis of these data using high-speed calculators. Static errors of GTE model realization by gas temperature are typical for this method [3].

II. A REVIEW OF EXISTING METHODS OF MEASURING GTE HIGH TEMPERATURES

To date, tungsten-rhenium thermocouples [4] BP 5/20 are widely used to measure high temperatures (more than 2000 °C), allowing to control the temperature up to 2500 °C.

When measuring the signal at the level of microvolts, interference from electric and magnetic fields [5] becomes a problem, which is solved by twisting a pair of thermocouple wires or using a shielded cable or laying wires in a metal tray, which is a protective screen.

The measuring device must provide signal filtering either at the hardware or software level, with intensive suppression of the network frequency and its harmonics. Figure 1 presents the scheme of measurement of GTE temperature by thermocouples.

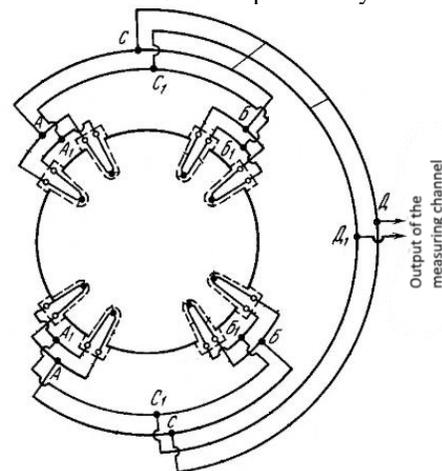


Fig. 1. The scheme of measurement of GTE temperature by thermocouples

When using platinum thermistors [6] Pt (385), which have the temperature measurement range minus 200 °C to 850 °C, it is necessary to use a variety of three-wire or four-wire bridge circuits, which greatly complicates the circuit when measuring the temperature field with a large number of thermistors. Figure 2 shows examples of (a – two-wire, b – three-wire, c – four-wire) bridge circuits for measuring temperature at one point of the controlled object.

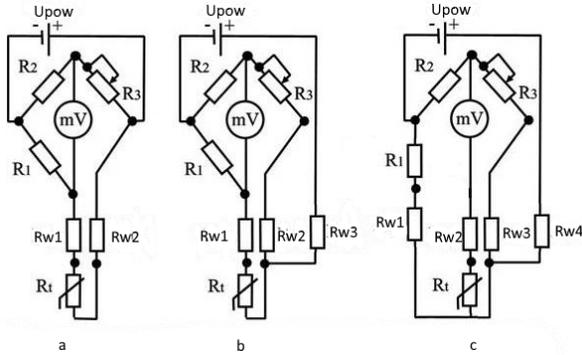


Fig. 2. The schemes of inclusion of a thermistor in the bridge circuit

III. PURPOSE THE STATEMENT OF THE RESEARCH PROBLEM

The creation of intelligent temperature sensors, in which are integrated functional converters and microcontrollers, makes it possible to process information in the sensor by a certain algorithm. In addition, it is possible to introduce correction, linearization of the transmission characteristics of the measured values in digital form. As a result, their actual metrological characteristics are significantly higher than the characteristics of traditional sensors. It contributes to the development and implementation of advanced gas turbine engine temperature measurement methods that require significant computational processing implemented in the sensor's microcontroller.

Based on these methods, sensors allow:

- to simplify the design requirements of the measurement object, which extends the use of sensors in different places and reduces the cost of their installation;
- to use new methods and principles of measurement that require quite complex computational processing of output signals;
- to make corrections for changes in resistivity and temperature coefficient of resistance from temperature;
- to process the information and give the current values of the measured value in the specified units.

The aim of the study is to improve performance and noise immunity, reduce computing operations and simplify the design of the sensor.

The objective of the study is to create a new method for continuous measurement of GTE gas temperature using the developed design of the thermistor and making automatic corrections to the coefficients depending on the temperature of the gases, providing high accuracy of measurement and output of current parameters in units of temperature.

IV. DEVELOPMENT OF A METHOD FOR MEASURING THE AVERAGE GTE GAS TEMPERATURE

Particular interest is the use of high-temperature tungsten-rhenium thermistors from tungsten with rhenium content

of 20%, which have [7] high sensitivity. They can be used for multiple up to 2300 and short-term [8] up to 2800 temperature field measurements.

To simplify the process of measuring high temperatures, it is proposed to use a multi-link RC-generator [9] with the required number of thermistors installed on the jacks of thermocouples.

The design of tungsten-rhenium thermistor using thermocouple manufacturing technology [4] is shown in figure 3.

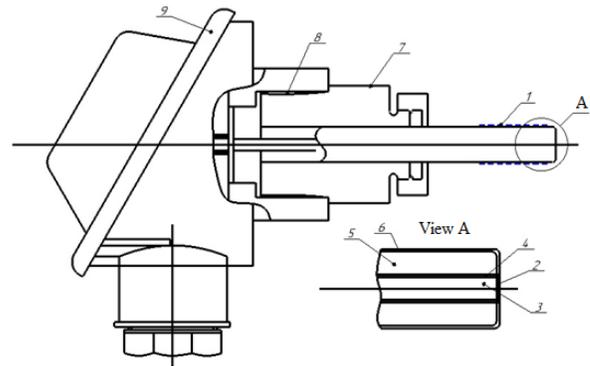


Fig. 3. The design of tungsten rhenium thermistor

Tungsten-rhenium thermistor is a thermal converter with a sensitive element 3 of tungsten-rhenium wire TR-20, placed in a housing 6 with a sealed cover 5 filled with inert gas (argon), providing protection of the thermistor. 2 – electron beam welding, 4 – electrical insulation, 7 – threaded fitting, 8 – glue filling, 9 – terminal box.

Molybdenum cover 1 is covered with a $MoSi_2$ and has a working life of up to 1000 hours.

Consider, as an example, the calculation of the parameters of the phasing RC-chain (PC) of the system for measuring the average temperature of GTE gas with the required number of thermistors, for example eight (Fig. 5), according to the method [9]. To meet the conditions of generation of the RC generator it is necessary to perform the phase balance and the amplitude balance. In accordance with table 1, we define the real part [10] of the transformation function (TF) of the PC, which is equal to $ReK_{16} = 21,55$. Therefore, the amplifier gain should be greater than 21.55.

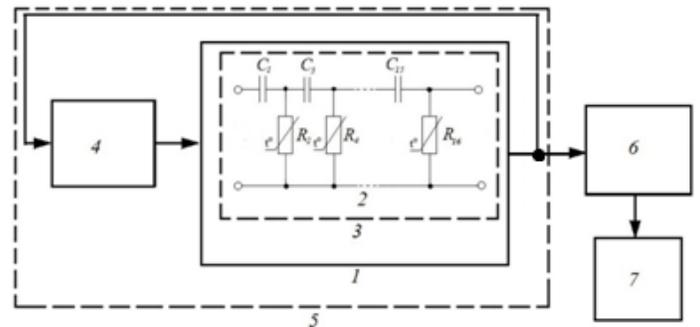


Fig. 4. System for measuring the temperature of the GTE gases

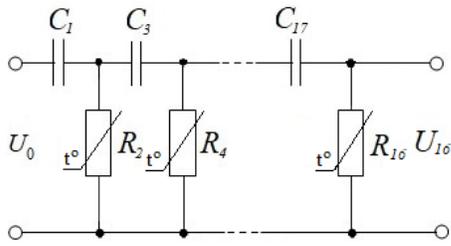


Fig. 5. Phasing RC-chain with thermistors

The new method of measuring the high temperature [11] of the inhomogeneous environment at object 1 is carried out as follows. Tungsten-rhenium thermistors 2 (temperature sensors) are evenly placed in a controlled environment, connected to external capacitors to form a phasing chain 3, which together with the amplifier 4 forms a setting generator 5 connected through [12] a microcontroller 6 with a digital indicator 7. With changes in the temperature of the controlled environment on the object, the resistance values of the thermistors forming the PC 3 of the generator 5 change. In accordance with the values of these resistances, the frequency of the generator 5 is set, which is converted by the microcontroller 6 frequency-code into units of the measured temperature and displayed on the indicator 7.

The quasi-resonance frequency of sixteen-shoulders RC-circuit is determined from the imaginary part of the TF when it is turned to zero:

$$f_0 = \frac{1}{2\pi k_{16} RC'} \quad (1)$$

where the coefficient k_{16} is determined from the expression:

$$\sum_{i=0,1,\dots}^P (-1)^i k_n^{2i+1} C_{0,5n+1+2i}^{2+4i} = 0,$$

where $C_{0,5n+1+2i}^{2+4i}$ – combinatoinis from $2+4i$ elements to $0,5n+1+2i$ element;

$$p = 0,25n - 1 \text{ – for even } 0,5n;$$

$$p = 0,25(n+2) - 1 \text{ – for odd } 0,5n \text{ PC.}$$

For the considered example of sixteen shoulders of PC $k_{16} = 0,286$ can be found also from the table 1.

TABLE 1. VALUE OF COEFFICIENT K_N AND ReK_N FROM THE NUMBER OF SHOULDERS N PC

n	6	8	10	12
k_n	2,449	1,196	0,739	0,590
ReK_n	29	24,70	23,46	22,77
n	14	16	18	20
k_n	0,373	0,286	0,227	0,185
ReK_n	22,26	21,55	20,58	20,11

Quasiresonance frequency of sixteen-shoulders RC-circuit was calculated in the program Deltafoc [13]:

$$f_0 = f_{pe3} = 1253,8 \text{ Hz.}$$

To create a small-sized thermistor, choose a wire diameter of 0.1 mm, as the smallest produced by the industry. The length of the wire is limited by the dimensions of the seat and is 80 mm

[15] as, for example, for a thermocouple type T80T. Wire resistance at 20 °C is equal to:

$$R = \frac{\rho l}{S} = \frac{1,93 \text{ Ohm} \cdot \text{mm}^2/\text{m} \cdot 0,08 \text{ m}}{7,8 \cdot 10^{-3} \text{ mm}^2} = 20 \text{ Ohm,}$$

where $\rho = 1,93 \text{ Ohm} \cdot \frac{\text{mm}^2}{\text{m}}$ from [16];
 $S = 7,85 \cdot 10^{-3} \text{ mm}^2$.

The working resistance of the wire R_t is calculated by the formula:

$$R_t = R_0(1 + \alpha_i t), \quad (2)$$

where are the temperature coefficient values α_i and the resistivity of the wire ρ_i also depend on the temperature.

Then the expression (2) takes the form:

$$R_t = \rho_i \frac{l}{S} (1 + \alpha_i t). \quad (3)$$

Table 2 presents the values of the specific resistance of the wire ρ_i , the temperature coefficient α_i and the values R_0 and R_t from temperature t .

TABLE 2. ELECTRICAL PARAMETERS OF TUNGSTEN-RHENIUM WIRE DEPENDING ON TEMPERATURE ρ_i

$t, ^\circ\text{C}$	ρ_i	α_i	R_0, Ohm	R_t, Ohm
20	1,93	0	19,7	19,7
100	2,54	0,00395	25,9	36,1
300	4,0	0,00383	40,8	87,6
500	5,2	0,00358	53,0	147,8
700	6,3	0,00333	64,2	213,8
900	7,25	0,00313	73,9	282,0
1100	8,05	0,00294	82,0	347,3
1300	8,7	0,00278	88,7	409,1
1500	9,3	0,00258	94,8	461,6
1700	9,85	0,00244	100,4	516,8
1900	10,3	0,00231	105,0	565,7
2100	10,65	0,00217	108,5	603,1
2300	10,9	0,00204	111,1	632,3

Data from table 2 are entered into the microcontroller and participate in the algorithm for calculating the temperature of the measured frequency of the generator.

Select the average generation frequency equal to 150 kHz, as optimal for the transmission of information along the communication lines, and calculate from (1) the capacitance of the capacitor C.

$$C = \frac{1}{2\pi f k_{16} R_t} = \frac{1}{2 \cdot 3,14 \cdot 0,286 \cdot 150 \cdot 10^3 \cdot 603} = 6800 \text{ pF,}$$

where $R_t = 632 \text{ Ohm}$ at $t = 2300 ^\circ\text{C}$.

The expression (1) for the smart sensor will then take the form:

$$f_0 = \frac{1}{2\pi k_{16} R_t C}. \quad (4)$$

The dependence of the frequency of the generator f on the temperature t of the tungsten-rhenium thermistors is shown in figure 6.

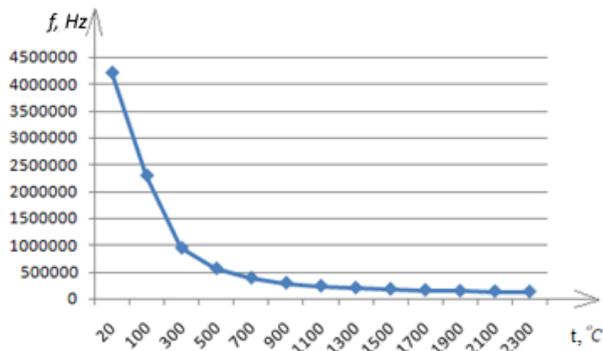


Fig. 6. The dependence of the oscillator frequency on temperature

CONCLUSION

The intelligent gas temperature measurement system GTE, based on tungsten alloy thermistors with rhenium TR-20, allows:

- 1 Continuously measure the average temperature up to 2300 °C using a two-wire communication line, which provides high noise immunity, because the informative parameter has a frequency nature.
- 2 Provides increased performance by 30% compared to thermocouples due to lower weight.
- 3 Reliability and efficiency of the method due to the small number of low-cost components used.

REFERENCES

- [1] Gulin A. I., Sukhinets Zh. A. *Analysis and synthesis of chain structures by the method of transformation functions*. Deutschland, Saarbrücken, LAP LAMBERT Academic Publishing, 2011, 199 p.
- [2] Petunin V. I., Sibagatullin R. R., Frid A. I. Error-correcting self-adjusting meter of the gas temperature of GTE. *Vestnik UGATU* [News of USATU], 2015. Vol. 19, № 1 (67), pp. 167-175.
- [3] Petunin V. I. Determination of gas temperature GTE using indirect measurements. *Izvestiya Vuzov. Aircraft equipment* [News of universities. Aircraft equipment], no. 1, pp. 51-55, 2008.
- [4] Ulanovskiy A. A. Experience in the use of wolframtones thermocouples BP5/20 in high-temperature thermometry. *Tetmoelektrichestvo* [Thermoelectricity], №1, 2009, pp. 86-92.
- [5] Matthew Duff, Joseph Tovey. Two methods of temperature measurement using thermocouples, *Modern electronics* №1, 2011, pp. 32-39.
- [6] State Standart 6651-2009. State system for ensuring the uniformity of measurements. Platinum, copper and nickel resistive temperature transducers. General technical requirements and test methods. Moscow, Standartinform Publ., 2011, 30p.
- [7] Gurevich A. M. Danishevskii, S. K., Smirnova N. And. Ipatov, S. I., Konstantinov V. I., Pavlova E. I. *Termopary dlya izmereniya vysokih temperatur s primeneniem termoelementov na molibdenovoj ili volframovoj osnove* [Thermocouple for measuring high temperatures with the use of thermoelectrodes on molybdenum or tungsten-based]. AS USSR no. 108438, publ. in "Bulletin of inventions", №4, 1958.
- [8] Fedik I. I., Deniskin V. P., Nalivaev V. I., Konstantinov V. S., Parshin N. M. Problems of high temperature measurements of IFA NTR. *Pribory+Automatizaciya* [Devices + automation], 3, 2002, pp. 20-27.
- [9] Sukhinets Zh.A., Gulin A. I., Matveev D. S., Nadrshin A. S. *Sposob izmereniya vysokoy temperatury neodnorodnoy sredy* [The method of measuring high temperatures of an inhomogeneous medium] Patent RF, no. 2624410, 2017.
- [10] Fazylova G. S., Gulin A. I., Sukhinets A.V. Intelligent sensor average temperature of the medium with inhomogeneous temperature field. *Sbornik nauchnyh trudov VII Vserossijskoj nauchno-tehnicheskoy*

konferencii "Perspektivy avtomatizacii tekhnologicheskikh processov dobychi, transportirovki i pererabotki nefi i gaza" [Collection of scientific works of the VII all-Russian scientific-technical conference "Perspectives of automation of technological processes of production, transportation and processing of oil and gas"]. Ufa, 2018, vol. 1. pp.79-85.

- [11] Gulin, A. I. Design of the multi-link RC-oscillators. *Izvestiya vuzov. Priborostroenie*. [News of Universities "Instrument-Making"], 2012, vol. 56, no.3, pp. 14-18.
- [12] Gamanyuk D. V. Domestic microcontrollers of new generation. *Sovremennaya elektronika* [Modern microelectronics], 2010. No. 5, pp. 16-21.
- [13] Sukhinets Zh.A., D. F. Mudarisov, Hanykov I. R., Gulin A. I. *Issledovanie amplitudno-chastotnyh i fazo-chastotnyh harakteristik cepnyh RC-skhem* [Calculation of the frequency of quasiresonance and transmission coefficient of the multi-link RC-structures]. Patent RF, no. 2003611147, 2003.
- [14] Technical conditions 11-75. The alloy wire of tungsten-rhenium annealed graded for thermoelectrodes thermocouples.
- [15] *Studfiles. Fajlovyj arhiv studentov*. Available at: <https://studfiles.net/preview/2216418/page:9/> (accessed 7 March 2018).
- [16] Nikolsky, B. P. Reference Book in Chemistry. Vol. 5: Raw materials and products of the inorganic substances industry. Processes and devices. Corrosion. Leningrad: Khimia, Leningrad Branch. 888 p. (in Russian).

Application of Modern Microelectronic Technology in Marshalling Process of Railway Stations

Michael A. Gordon,
Chief specialist of the Institute "Giprotranssignalsvyaz" –
Department of JSC "Roszheldorproject"
Saint-Petersburg, Russia
gordon_ma@mail.ru

Alexey N. Kovkin,
PhD, Associate Professor at
"Automation and Remote Control on Railway" Department
Emperor Alexander I St. Petersburg State Transport University
Saint-Petersburg, Russia
akovkin@yandex.ru

Dmitry V. Sedykh,
Chief engineer of Group of Companies «IMSAT»,
engineer at "Automation and Remote Control on Railways"
Department,
Emperor Alexander I
St. Petersburg State Transport University
Saint-Petersburg, Russia
sedyhdmitriy@gmail.com

Anton A. Movshin
Head of group of the Institute "Giprotranssignalsvyaz" –
Department of JSC "Roszheldorproject"
Saint-Petersburg, Russia
movshinaa@rzd.ru

Oleg A. Abramov,
Senior Researcher of Computer Railway Technology Center at
"Automation and Remote Control on Railway" Department
Emperor Alexander I St. Petersburg State Transport University
Saint-Petersburg, Russia
olegabramov@mail.ru

Abstract — In the contemporary world of global digitalization, a problem arises concerning compliance of the marshalling process control systems with the modern requirements. To solve the problem, a Train Breaking-Up Control System "SURS GTSS" has been developed, which permits to control the hump automation and remote control devices using microelectronic technology. The system provides for safety of shunting at a gravity hump yard and improves quality of train breaking-up. At the same time, the operating expenses, engineering, construction, and commissioning periods and labor hours decrease, and relay equipment is almost missing. As a result, the preset hump capacity parameters are achieved, service and system troubleshooting quality improves. The article describes the system structure, provides a description of the system components with their functionality, and shows a block diagram of a hardware package of the Hump Microprocessor-Based Interlocking "GMC GTSS" subsystem. Remote control diagrams are also presented for electric switch machines, humping and shunting signals, hump retarders are presented. A structure of the asset controllers governing the devices is described.

Keywords — microelectronic control of assets, SURS GTSS, GMC GTSS, GAC-ARS GTSS, train breaking-up automation.

I. INTRODUCTION

Shunting stations are a central component of the railway complex to process the railcar traffic volumes and make up freight trains in an optimal mode so that the presence time of a railcar at the station is as short as possible and reasonable in process terms. Shunting stations consist mostly of a receiving yard, where the trains intended for further splitting arrive, a gravity hump yard, a marshalling (hump) yard, and a departure yard, wherefrom the made-up trains depart further. A gravity hump yard is a man-made hummock to let down the railcars (cuts) uncoupled on a marshalling yard track under gravity [1- 4].

To achieve as much efficient freight train making-up as possible, the gravity hump yards of the world employ both mechanical and automatic equipment. The contemporary train

shunting automation systems consist of two parts – hump automatic interlocking and automatic control of uncoupling speed system. In Russia, the initial hump automatic interlockings appeared in 1946 and automatic control of uncoupling speed systems in 1961. The systems were relay-controlled. At present, the railways of the world introduce only microprocessor-based systems, including the hump ones. Germany has developed such shunting station control systems as "ADRS", "BLR", "ITS", and "MSR-32"; France – "Saxbi"; USA – "ATGS", "SPM", "DDC – III"; and Russia – "KSAU SP", "GAC-ARS GTSS", "SKA-SP". However, the integrated microprocessor-based hump systems use a relay-operated interface to control the trackside assets (color light signals, electric switch machines, railcar retarders, etc.) At present, a task arises to shift from the relay-operated towards microelectronic equipment [5 - 8].

II. TRAIN BREAKING-UP CONTROL SYSTEM "SURS GTSS"

In order to reject actually the electromagnetic relays, a comprehensive Train Breaking-Up Control System "SURS GTSS" has been developed, which permits to control the hump automation and remote control devices employing microelectronic technology, without using both large-size relays as well as micro ones [9 - 10].

A microprocessor-based system "SURS GTSS" consists of the following subsystems:

- "GMC GTSS" – a safe control system of the trackside assets (railway switches and color-light signals) and safe interaction with the electric interlocking devices to draft in, hump, and break up railway trains;
- "GAC GTSS" – an automatic path point setting subsystem;
- "ARS GTSS" – a displayed speed subsystem (retarder control).

Depending on the requirements for an asset automation extent and the asset specifications (gravity hump yard capacity), the system implementation options are possible, which offer a different functional set and, as a consequence, hardware and software implementation:

- cut path control process automation as a separate subsystem;
- gravity hump yard cut rolling-down speed control process automation as a separate subsystem;
- integrated implementation of path control and cut rolling-down speed control processes within the structure of one system.

Under any implementation option, the system integrates the control and troubleshooting and recording functions of the system and controlled asset [11, 12].

III. SUBSYSTEM “GAC-ARS GTSS”

The central purpose of an “GAC-ARS GTSS” subsystem is:

- implementation of an estimated gravity hump yard capacity;
- automatic accumulation and implementation of the paths in the course of automatic railcar cut shunting;
- implementation of a preset speed for railcar exiting from the interval retarder position;
- target braking of railcar cuts at the yard retarder position;
- quality improvement of train splitting;
- running efficiency improvement of the hump yard and operating staff;
- working environment and safety improvement;
- splitting process control operability improvement.

The system has been designed on a modular basis using dedicated and standard equipment and network data interactions.

An “GAC-ARS GTSS” subsystem consists of the following modules to implement the following functions:

- computer complex control cabinet with galvanic isolation modules – data input and processing, railcar cut movement monitoring;
- displayed speed subsystem cabinet with local controller modules – automatic railcar cut speed control in the interval and yard retarder zones;
- process supervision equipment – railcar cut rolling-down path presetting and railcar cut performance indication;
- automated workstation of Hump Electrician – status displaying of trackside assets, railcar cut rolling-down process, input signal logging, automatic shunting list recording and printing;
- automated workstation of Hump Foreman – displaying of the controlled and monitored assets and generation of the control actions;

- automated workstation of Hump Operator – displaying of a railcar cut retarding process at the retarder position, displaying of the track occupancy monitoring device indications.
- gateway – functions of communication with adjacent systems.

The supporting equipment of the system is both tower and trackside assets of the mechanical equipment for the gravity hump yard [13, 14].

Ultimately, the system supports communication with any required adjacent system implemented in the network structure. Communication is through a dedicated gateway.

IV. SUBSYSTEM “GMC GTSS”

The central purpose of a subsystem “GMC GTSS” is:

- safe interaction with the electric interlocking devices of the receiving park to draft in, hump, and break up trains;
- safe control of the railway switches and color-light signals with a check of key dependencies;
- control of the retarder position equipment to provide for required intervals between the railcar cuts;
- reduction of the tower equipment requirement;
- improvement of the gravity hump yard and operating staff performance;
- working improvement and safety improvement;
- splitting process control operability improvement.

A “GMC GTSS” subsystem consists of the following modules implementing the following functions:

- automated workstation of Hump Foreman – displaying of the controlled and monitored assets and generation of the control actions;
- data input/output (communication) cabinet from the digital devices (Cabinet-I) – control and monitoring of the assets offering an digital interface. The following functions are implemented by means of the equipment:
 - a) control of the railcar quantity indicator;*
 - b) control of the breaking-up speed indicator;*
 - c) input of the retarder status and operation data;*
 - d) issue of the braking/release commands to the railcar retarders;*
 - e) input of the data from the weight measuring device;*
 - f) input of the data from the environment monitoring devices;*
 - g) input of the data from the rolling stock movement speed transducers.*
- data input/output cabinet from the discrete relay devices (Cabinet-RD) – control and monitoring of the assets offering a discrete relay interface. The following functions are implemented by means of the equipment:
 - a) monitoring of the track circuits in the electric interlocking adjacent zones;*
 - b) monitoring of the axle counter status;*
 - c) input of the data on marshalling track protection;*

- d) *input of an emergency hump light signal turning-off command;*
- e) *annunciation and signaling in the breaking-up zone;*
- f) *issue of the control actions to the facility cleaning devices.*

- central processor cabinet of hump microprocessor-based interlocking with two redundant sets of central hump interlocking processors – implementation of the central dependence subsystem functions;
- railway switch and light signal control cabinet – control and monitoring of the light signals and electric switch machines;
- set switching device installed in the central processor cabinet – switching between the operating and standby sets;
- automated workstation of Hump Electrician consisting of a PC system unit, monitor, keyboard, and printer – status monitoring of the trackside assets, monitoring and analysis of troubleshooting data on the system operation, system performance recording, logging and printing of the generated records;
- network equipment consisting of two redundant network hubs installed in the Central Processor Cabinet – support of interactions among the central processor workstations, automated workstations, other system cabinet workstations, hardware and software packages;
- server equipment installed in a service cabinet (Cabinet-D) – logging and archiving into a database, safe storage of the system records, archives, and installation software;
- gateway equipment installed in a service cabinet (Cabinet-D) – communication with adjacent systems such as an automated shunting station control system, a hump engine control system, a monitoring system [15 - 17].

A hardware package structure of the “GMC GTSS” subsystem is shown in Fig. 1.

V. RAILWAY SWITCH CONTROL

To control the electric switch machines as part of “GMC GTSS”, safe interfacing controllers “KBS-SG” are used.

The switch positions are monitored through signal pickup from the proximity sensors being part of humping electric switch machines. Interactions between the power and controller modules are by means of dynamic and potential TTL-level signals transmitted via the parallel bus bars implemented on a “KBS-SG” motherboard. Circuit units switching the points to «plus» and «minus» are provided with independent power ports connected to external sources via separate protection devices .

The “KBS-SG” controllers provide for the following:

- reception and implementation of commands duplicate the RS-485 interface from the central processor cabinet of the hump microprocessor-based interlocking as well as ensures the collection of the checking information

and the transmission through the interface into the central processor cabinet;

- implementation of control signals by the power modules;
- pickup of control signals from the power modules;
- monitoring of signal identity within the redundant structure and implementation of the lower-level logic dependencies requiring a high speed of performance. Such logical dependencies include inhibition of point switching in case of an equipment failure, abortion of point switching as control is gained, automatic return of the point upon a failure to gain control.

The “KBS-SG” has been designed under a 1oo2D architecture due to availability of two controller modules. This architecture is allowed for application on Russian Railways. In case one of the controller modules fails in the course of switching, the initiated point switching will be completed as the controller functions are performed simultaneously by the both controller modules in the course of point switching.

A “KSB-SG” module structure and a point machine connection diagram are shown in Fig. 2.

The power modules are responsible for direct interaction with the humping electric switch machines, pickup and logical processing of the dynamic control signals from the controller modules designed under a 1oo2D architecture, generation of the dynamic signals indicating the switch position and the switch section status. In addition, the power modules generate signals for the controller modules, which indicate operability of a power module both in the course of point switching as well as during the time intervals between switching actions. A specific feature of the hump KBS power modules is availability of the independent functional units for point switching in different directions, which permits to return the point back in case of a single power module failure. A power module is capable of commuting a current up to 8A (friction current) and allows for short-time flowing of a current up to 30A (engine starting current). The operation mode of a power module is an intermittent cycle. Two double-wire inputs have been provided for monitoring of the point status, where a rectified voltage is supplied from the signal winding of the point operating gear proximity transducers. Every power module contains two switching and position monitoring units and an operability and vacancy monitoring unit.

A switching and position monitoring unit consists of a safe logic cell designed under a 1oo2D architecture, a control signal generator, a power converter, a position monitoring pulse generator. Input signals arrive at the safe logic cell through optical isolator components from the controller modules. An output signal of the safe logic cell is supplied to the control signal generator as a power supply voltage. The control signal generator governs the converter power keys. The power converter generates a constant voltage of 220 V for operation of the point motor. The power converter includes a bridge inverter on IGBT transistors, a rectifier bridge, a dividing transformer, an isolating diode, and a smoothing capacitor. The position monitoring pulse generator has been designed as a voltage-controlled relaxation generator to generate signals for the controller modules under a full shift point position.

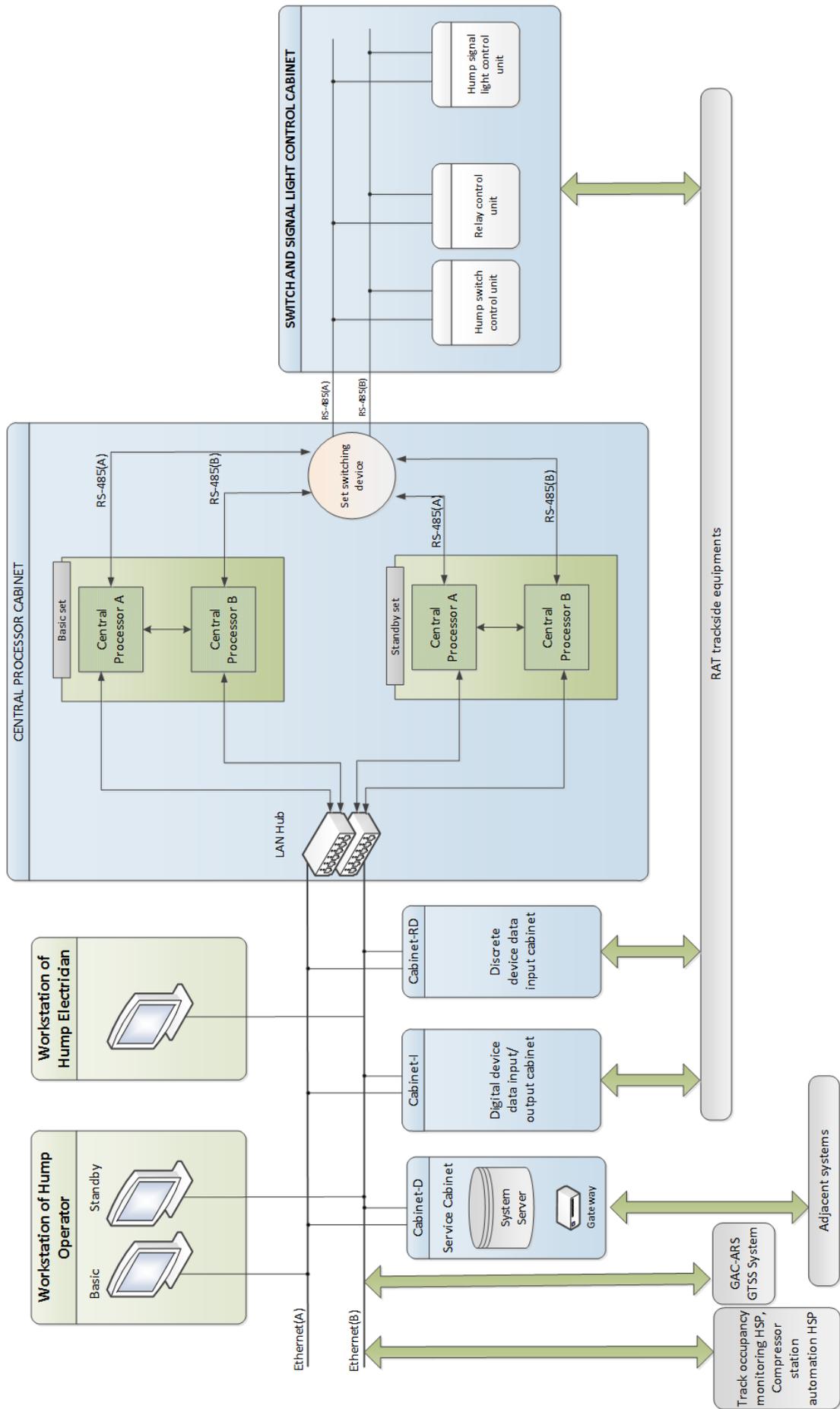


Fig. 1. Hardware package structure of "GMC GTSS" subsystem

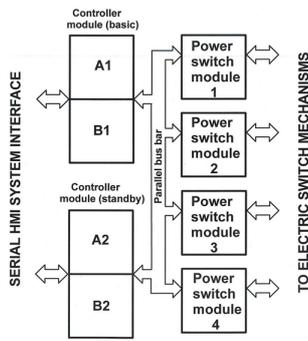


Fig. 2. "KSB-SG" module structure and point machine connection diagram

An operability and vacancy monitoring unit consists of an operability monitoring circuit and a vacancy monitoring pulse generator. The operability monitoring circuit has been designed to monitor integrity of the power module and actuating circuit of the power switch machine, monitor the thermal mode of the module, availability of the required power supply voltages. The vacancy monitoring pulse generator is the same as the position monitoring pulse generator [18].

A functional switch control power module diagram is shown in Fig. 3.

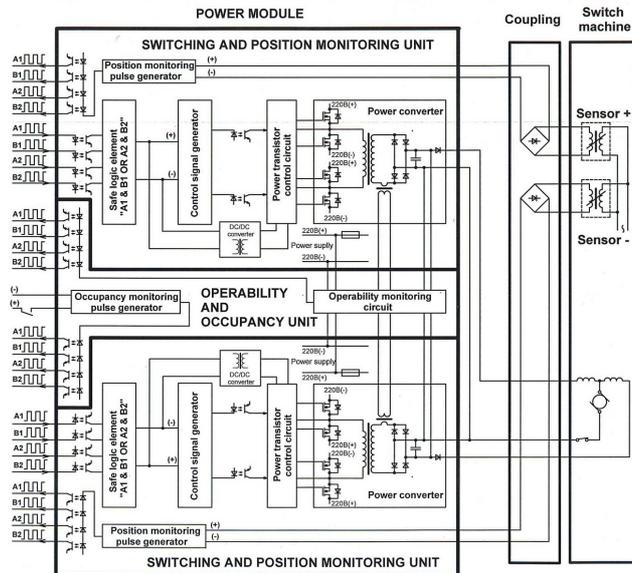


Fig. 3. Functional switch control power module diagram

VI. COLOR LIGHT SIGNAL CONTROL

To control the color light signals, safe interfacing controllers "KBS-SVG" and "KBS-SVM" are used as part of "GMC GTSS", which are designed for humping and shunting light signals, respectively.

The light signal lamps are controlled via a double-wire circuit. The lamps are connected through signal transformers. Interactions between the power and controller modules are by means of TTL-level dynamic signals transmitted via parallel bus bars implemented on a light signal KBS motherboard.

The "KBS-SVG" and "KBS-SVM" controllers provide for the following:

- reception and implementation of commands duplicate the RS-485 interface from the central processor cabinet of the hump microprocessor-based interlocking as well as ensures the collection of the checking information

and the transmission through the interface into the central processor cabinet;

- implementation of control signals by the power modules;
- pickup of control signals from the power modules;
- monitoring of signal identity within the redundant structure;
- implementation of the lower-level logic dependencies according to the light signal type.

The power modules are responsible for control of the light signal lamps connected through the signal transformers, pickup and logic processing of dynamic control signals from the controller modules designed under a 2oo2 architecture, generation of the dynamic signals indicating the current value on the output circuit and integrity of the cold lamp filaments. The output voltage of the power modules is of a rectangular shape [19].

A "KBS-SVG" module structure are shown on Fig. 4.

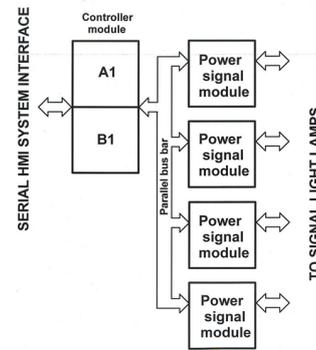


Fig. 4. "KBS-SVG" module structure

VII. RAILCAR RETARDER CONTROL

To control the railcar retarders, air collectors with renovated control equipment are used, which apply a principle of dividing the compressed air flows into a working and controlling flow.

Brake activation / release commands to the actuators as well as the operating air collector pressure, current command, target control equipment performance indicator values are transmitted via a redundant interface link RS-485. A retarder control diagram is shown in Figure 5.

The manual mode can be applicable to develop automatically simulators of real train traffic control systems based only on their technical description as well as to monitor correctness of circuit designs for automatic examination of the circuit designs of train control systems [20].

VIII. CONCLUSION

Operation of microelectronic intelligent technology for control and monitoring of the trackside assets, including that on railway humps, supports digitalization of a railcar classification process, which permits to undertake real-time troubleshooting and monitoring of the devices. The developed by authors comprehensive "SURS GTSS" system permits to improve the digitalization standard of the marshalling yard operations. The system using microelectronic technology has permitted to reject totally relay-based integration with the controlled sidetrack assets. The system provides for safety of the hump railcar breaking-up and improves quality of the train

splitting. At the same time, the operating expenses, the engineering, construction, and commissioning periods and labor hours decrease, and relay equipment is almost missing. As a result, the preset hump capacity parameters are achieved, service and system troubleshooting quality improves. Rolling-down cut speed control computation algorithms are expected to be developed further. This task becomes relevant now to support classification of the railcars containing dangerous goods by means of a gravity hump yard [21 - 23].

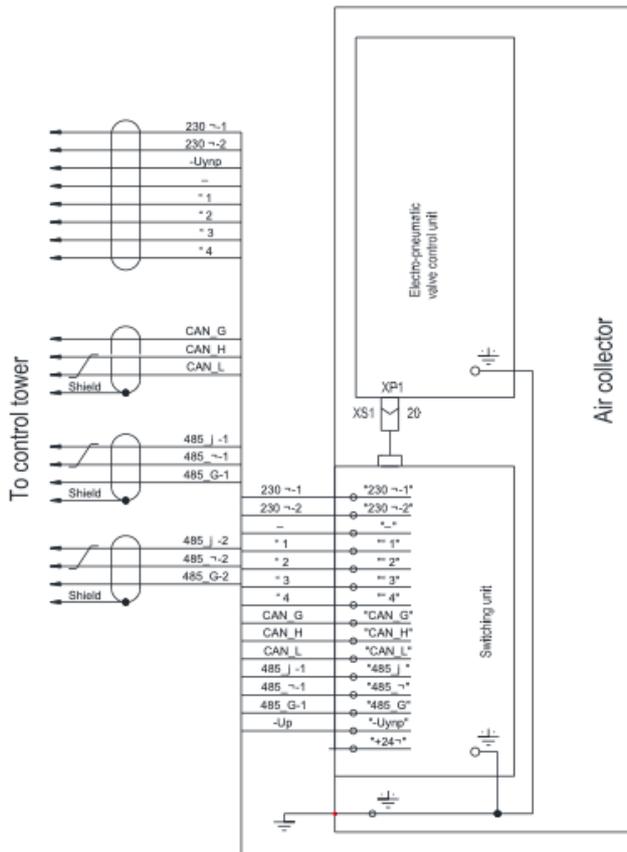


Fig. 5. Retarder control diagram

REFERENCES

[1] S. Zářecký, J. Grůň, and J. Žilka, "The newest trends in marshalling yards automation," *Transport problems*, 2008, Tom 3, Volume 4, Part 1, pp. 87-95.

[2] T. Teixeira da Mota, "Evaluation of the operational gains obtained from the automation of a marshalling yard" [online] Available: https://fenix.tecnico.ulisboa.pt/downloadFile/395144992791/Abstract_13Dez.pdf.

[3] L. Dimitrov, S. Purgic, P. Tomov, and M. Todorova, "Approach for Development of Real-Time Marshalling Yard Management System," 2018 International Conference on High Technology for Sustainable Development (HiTech), Sofia, 2018, pp. 1-5. doi: 10.1109/HiTech.2018.8566369.

[4] G. Theeg, S. Vlasenko (eds.) "Railway Signalling & Interlocking: International Compendium", 2nd revised edition, PMC Media House GmbH, 2017, 458 p.

[5] C. Li, X. You, X. Liu, and P. Wu, "Study on the Method of Complex Lines Selection and Automated Processing for Railway Marshalling Yard," International Conference on Energy, Power and Electrical Engineering (EPEE 2016), pp. 306-310.

[6] "Shift2Rail Multi Annual Action Plan", Brussels, 2015, [online] Available: www.shift2rail.org/wp-content/uploads/2013/07/MAAP-final_final.pdf.

[7] "Identification of relevant information about train classification process and marshalling yard sorting methods", Report for Deliverable D4.1. of EU-SMART-project under GA-No. 730836, 2017, [online] Available: <http://smartrail-automation-project.net/index.php/results/deliverables>.

[8] N.A. Nikiforov, "System of automatization of hump processes" (in Russ.), *Automation, Communication and Informatics*, 2006, issue 9, pp. 39-40.

[9] O. Strelko, H. Kyrychenko, Y. Berdnychenko, and S. Hurinchuk, "Automation of Work Processes at Ukrainian Sorting Stations," *International Journal of Engineering & Technology*, 2018, issue 7 (2.23), pp. 516-518.

[10] V.V. Khóroshev, D.V. Efanov, and G.V. Osadchii, "Ways of Development of Periodical and Continuous Monitoring Means for Automatic Devices on Marshaling Yards," *Proceedings of 1th International Russian Automation Conference (RusAutoCon)*, Sochi, Russia, September 9-16, 2018, pp. 1-5, doi: 10.1109/RUSAUTOCON.2018.8501720.

[11] S. Gestrelus, F. Dahms, and M. Bohlin, "Optimisation of simultaneous train formation and car sorting at marshalling yards," *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen)*, 2013.

[12] S. Gestrelus, "Mathematical models for optimising decision support systems in the railway industry," PhD Thesis Mälardalen University Licentiate Thesis No. 196, 2015, ISSN 1651-9256.

[13] "Overall framework architecture and list of requirements for real-time marshalling yard management system," Report for Deliverable D4.2. of EU-SMART-project under GA-No. 730836, 2017, [online] Available: <http://smartrail-automation-project.net/index.php/results/deliverables>.

[14] H. Djellab, C. Mocquillon, "An efficient heuristic method for the hump yard management problem," *Proceedings of 12th World Conference on Transport Research*, 2010.

[15] P.S. Rakul, I.N. Zhmudanov, and N.A. Nikiforov, "Hump microprocessing interlocking GMC GTSS" (in Russ.), *Automation, communication and Informatics*, 2019, issue 1, pp. 27-30.

[16] I.N. Zhmudanov, and M.A. Gordon, "Hump microprocessing interlocking is base of safety for disbanding of trains (GMC GTSS)" (in Russ.), *Proceedings of XIX All-Russian scientific and practical conference "Safety of train movements"*, Moscow, Russia, November 8-9, 2018.

[17] "Description of automation/optimization requirements and capabilities of decision making process in Marshalling yards and Terminals," Report for Deliverable D2.1. of EU-ARCC-project under Contract No. H2020 - 730813/MC S2R-CFM-IP5-02-2015, 2017.

[18] A.B. Nikitin, and A.N. Kovkin, "Point machines control in computer systems of hump interlocking" (in Russ.), *Automation on Transport*, 2015, vol. 1, issue 1, pp. 51-62.

[19] A.B. Nikitin, A.N. Kovkin, and A.D. Manakov, "Using the small-sized power relays for design of safe interface devices within the computer systems of railway automation" (in Russ.), *Automation on Transport*, 2018, vol. 4, issue 2, pp. 264-278.

[20] J. Adlbrecht, B. Hüttler, N. Ilo, and M. Gronalt, "Train routing in shunting yards using Answer Set Programming," *Expert Systems with Applications*, 2015, vol. 42, no. 21, pp. 7292-7302.

[21] E. Dahlhaus, P. Horak, M. Miller, and J. Ryan, "The train marshalling problem," *Discrete Applied Mathematics Volume*, Elsevier, 2000, vol. 103, pp. 41-54.

[22] N. Boysen, M. Fliedner, F. Jaehn, and E. Pesch, "Shunting yard operations: Theoretical aspects and applications," *European Journal of Operational Research*, pp. 1-14, 2012.

[23] D.V. Efanov, G.V. Osadchy, and V.V. Khoroshev, "Testing of Optical Sensors in Measuring Systems on Railway Marshalling Yard," *Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS'2018)*, Kazan, Russia, September 14-17, 2018, pp. 225-230, doi: 10.1109/EWDTS.2018.8524798.

Ternary Parity Codes: Features

Dmitrii V. Efanov,
DSc, Professor at "Automation, Remote Control and Communication on Railway Transport",
Russian University of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

Abstract—For organizing offline and online diagnostics systems, redundant coding is widely used. Codes are focused on error detection in the codeword bits. At present, the automation devices function in binary logic and consequently binary redundant codes are used in the development of technical diagnostic tools. Nevertheless, several specialists note the ternary number system and automation devices advantage that implement its principles. This work is devoted to the simplest ternary codes – ternary parity codes. Some ternary parity codes properties are described, and the need for a more detailed account of the error type in the ternary redundant code's data vectors is noted. It is proved that the described ternary parity code, built based on using a single convolution function modulo three, has the data vectors uniform distribution property between check single-trit vectors. In addition, was introduced the concept of code with the least total number of undetectable errors in the ternary code data vectors for the lengths fixed values of data and check vectors. It is proved that such a code has an data vectors uniform distribution between all possible check vectors. Formulas for calculating the total number of undetectable errors in such codes are given.

Keywords—automation devices based on ternary logic; automation devices; ternary logic; digital devices offline and online control; ternary parity code; error checking at the device outputs; error types.

I. INTRODUCTION

In the modern world, automation and computing devices are implemented based on binary logic. This contributes to the ease of implementation and operation relative reliability. The recognition of this number system for the computing systems implementation was recorded in the well-known von Neumann architecture [1]. However, the ternary logic (ternary logic or three-valued logic) using can be very promising. The ternary logic advantages are explained by a denser number recording. In addition, the ternary logic "covers" the binary and can use all its advantages, and the computing devices implemented on the ternary logic should be faster.

Over the years computer technology development has been successful attempts to implement ternary computing systems. For examples are the "Setun" system, implemented at Moscow State University in 1958 (under the direction of N. P. Brusentsov) [2], and TCA2 (version v2.0), created at the California Polytechnic University in 2008 (J. Connelly, K. Patel, A. Chavez) [3]. The whole direction of scientific research relates to the ternary logic devices development on the elements of binary logic. Such devices, for example, are described in [4 – 6]. Some quantum computers developers speaking about the ternary logic advantages over binary, proposing to use quotes instead of qubits, explaining this by a serious decrease in the number of quantum gates [7].

In addition, the ternary logic can be effective for the data protection of modern devices using IoT [8].

The upcoming fourth industrial revolution, associated with tremendous computer technology development and the cyber-physical control systems development. That raises the researcher's interest in the automation functioning devices principles in the ternary logic [9 – 14]. With the implementation of such devices, albeit more often, as some models, the methods of their technical diagnostics are being developed [15, 16].

One of the well-known approaches to building reliable fault-tolerant automation systems and devices with fault detection is the use of redundant coding. Error-detection codes are widely used as codes that have less redundancy than error-correcting codes and do not allow to accumulate faults until a complete system failure [17 – 19]. This paper focuses on elementary ternary codes – ternary parity codes. Such codes can be built in various ways. For example, one check bit can be used, in which the convolution modulo three value of all data bits are written, or two check bits, each of which is intended for a separate the parity sum checking of the data vector significant numbers [20].

Consider the ternary parity codes features in relation to the digital systems offline and online diagnostics tasks.

II. TERNARY PARITY CODES

In the ternary notation, there are three logical signals. In various sources they are denoted differently, for example (0, 1, 2), (–1, 0, 1), (0, ½, 1) and others. However, the ternary number system is implemented in two versions, forming an asymmetric and symmetric number systems. In the asymmetric number system, the notation for logical signals (0, 1, 2) is used. In a symmetric number system, the symbols for the notation for logical signals are as follows: (–1, 0, 1). Symbols denoting logic levels for the ternary logic automation can be used any but considering the precedence. Further, in the reasoning, we will use the following symbols set $x, f \in \{i; 0; 1\}$, focusing on a variant of the symmetric number system. It should be noted that all the results described below are universal and not tied to a specific type of notation system, and the third signal symbol in the «*i*» form is used for recording ternary vectors simplicity.

As mentioned earlier, one data trit can take values from the set $f \in \{i; 0; 1\}$, in other words, have 3 values options. If we are talking about some automation device with a known structure with the number of outputs – m , then at its outputs m -bit ternary vectors $\langle f_1 f_2 \dots f_{m-1} f_m \rangle$ (data vectors) is formed at any time during operation. Their total number is determined by the 3^m value. The device internal structure failures lead to the error's propagation on its outputs, causing simultaneous distortions of several dataal discharges at once. In other words, caus-

ing errors in the data vectors. The number of such errors is large and is determined by the data vector bits number:

$$N_m^3 = 3^m(3^m - 1). \quad (1)$$

In expression (1), the first factor is the value characterizing the power of the set of complete sets of vectors on which errors are considered. The second factor determines the power of the set of error vector combinations for each data vector. The number of variants of erroneous transitions for each data vector, determined by the $\sum_{d=1}^m C_m^d = 3^m - 1$ value for the ternary one.

For example, for $m=3$ there is a number $N_3^3 = 702$ (more than 12 times greater than for the binary parity code [21]). With an increase in the m value, the number of the ternary data vectors distortion variants increases significantly.

To control the occurrence of the error at the digital device's outputs, redundant codes are used. They may have different construction principles. It should be noted that the building ternary codes theory is actively developing [22 – 25], and methods for constructing block uniform ternary codes, for example, an even-weight code analogue from binary logic – a compositional ternary code [26, 27] are also known.

For solving technical diagnostics problems, including working diagnostics, if it is necessary to reduce device redundancy and simplify technical diagnostic tools, the simplest redundant codes can be useful. The simplest is the ternary parity code. It is constructed in the same way as a binary parity code by using the convolution modulo three functions, or logical addition modulo three:

$$g = f_1 \oplus f_2 \oplus \dots \oplus f_{m-1} \oplus f_m. \quad (2)$$

Thus, the ternary parity code has only one check bit, which allows detecting any single errors in the data vectors.

The addition function for modulo three is described using Table 1, where the input variables values are indicated in rows and columns, and the function values are recorded at the intersections of the rows and columns.

TABLE I. THE MODULO ADDITION FUNCTION $M=3$

x_1	x_2		
	i	0	1
i	1	i	0
0	i	0	1
1	0	1	i

It is necessary to note the considered function important property. On one-third of the input sets, it takes the i values, on the other third parts of the input sets of value 0 and on the remaining third of the input sets of value 1 . In other words, the distribution of the values is uniform over all input sets.

It should be noted that practical implementations of the addition modulo three functions on traditional microelectronic components are known. For example, in Fig. 1 is a schematic diagram implementation of the addition function modulo three from [28], implemented on binary logic elements. The circuit contains transistors VT_1 and VT_2 , which are a pair of mutually additional amplifiers. The signal at the output of this pair coincides in phase with the input signal and is twice as large in amplitude. The transistor VT_3 bias voltage is chosen so that the

signal through the resistances R_1 and R_2 switch the current flowing through the elements VD_1 and R_3 , which leads to an increase in the transistor VT_2 output current.

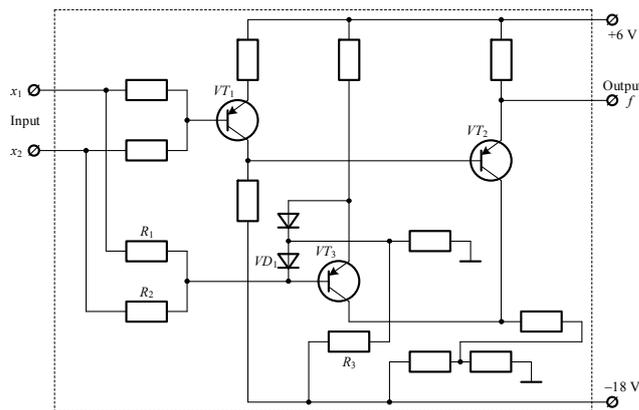


Fig. 1. Schematic diagram of the modulo addition function $M=3$.

Denote the ternary parity code as a $P(m,k)$ -code. For example, a code $(P(3,1)$ -code) is given in Table 2.

TABLE II. $P(3,1)$ -CODE VECTORS

No.	f_1	f_2	f_3	g
1	i	i	i	0
2	i	i	0	1
3	i	i	1	i
4	i	0	i	1
5	i	0	0	i
6	i	0	1	0
7	i	1	i	i
8	i	1	0	0
9	i	1	1	1
10	0	i	i	1
11	0	i	0	i
12	0	i	1	0
13	0	0	i	i
14	0	0	0	0
15	0	0	1	1
16	0	1	i	0
17	0	1	0	1
18	0	1	1	i
19	1	i	i	i
20	1	i	0	0
21	1	i	1	1
22	1	0	i	0
23	1	0	0	1
24	1	0	1	i
25	1	1	i	1
26	1	1	0	i
27	1	1	1	0

The error detection characteristics by separable codes are significantly affected by the number of data vectors corresponding to each check vector. The data vectors separation from their full set of a given length into subsets corresponding

to specific check vectors from their full set of a given length will be called the data vectors distribution between the check vectors. If the share of data vectors of their total number in the full set of vectors corresponding to each of the check vectors of the complete set is the same, then we will call this *distribution* of data vectors between the check vectors a *uniform distribution*.

It should be noted an important pattern inherent in $P(m, k)$ -codes.

Theorem 1. *The $P(m, k)$ -codes have an data vectors uniform distribution between all check vectors.*

The Theorem 1 proof arises from the fact that the modulo-three addition function used in calculating the check digit has a uniform distribution with respect to all input sets. When building the code, all input vectors are considered. **The theorem is proved.**

The Theorem 1 position in the logical devices operational diagnostics theory has an important role. A code that has the data vectors uniform distribution property among all possible check vectors will have the least number of undetectable errors among all possible codes for given values of m and k (k is the number of check bits).

Theorem 2. A ternary code with parameters m and k will have the minimum total number of undetectable errors, provided that all 3^m data vectors are equally distributed among all 3^k check vectors, and the total number of undetectable errors in such a code will be determined by the formula:

$$N_{m,k}^{min} = 3^m(3^{m-k} - 1). \quad (3)$$

The Theorem 2 proof. The ternary separate code is specified as a table (Table 3). To each of the 3^k check vectors, we assign the check group $i \in \{0, 1, \dots, 3^k\}$ to correspondence (in fact, the group number corresponds to the decimal representation of the binary number written in the check vector). Each check group will contain q_i data vectors.

TABLE III. DESIGNING A TERNARY SEPARATE CODE

Check vector				
0	1	...	$3^k - 1$	3^k
The number of data vectors				
$2C_{q_0}^2$	$2C_{q_1}^2$...	$2C_{q_{3^k-1}}^2$	$2C_{q_{3^k}}^2$

Since the undetectable error is the one that translates the data vector of one check group into the data vector of the same check group, the number of undetectable errors in each check group will be determined by twice the number of all possible transitions of each vector into each:

$$2C_{q_i}^2 = q_i(q_i - 1). \quad (4)$$

The total number of undetectable errors will be:

$$N_{m,k} = \sum_{i=0}^{2^k} 2C_{q_i}^2 = \sum_{i=0}^{2^k} q_i(q_i - 1). \quad (5)$$

If all 3^m data vectors are distributed evenly between all 3^k check vectors, that is, $q_0 = q_1 = \dots = q_{2^k} = q$, then each check group will contain $q = \frac{3^m}{3^k} = 3^{m-k}$ data vectors. From formula (4) it follows that the number of undetectable errors in each group will be determined by the value:

$$2C_q^2 = q(q - 1) = 3^{m-k}(3^{m-k} - 1). \quad (6)$$

Multiplying expressions (6) by the 3^k value, obtain formula (3).

Proving that it is precisely formula (3) that determines the undetectable errors total number minimum in data vectors for given parameters m and k .

Suppose that code with an uneven all data vectors distribution between all check vectors will not detect a smaller number of errors in the data vectors than a code with their uniform distribution. Since the total number of data vectors is invariably equal to 3^m , there will be more in some check groups and fewer in some check groups.

Suppose that data vectors $q = 3^{m-k}$ are present in one check group instead of $(q - b)$ data vectors, and b data vectors are distributed among all other check groups. In this case, the number of undetectable errors in the check group with a reduced number of data vectors will be:

$$(q - b)(q - b - 1) = (q - b)^2 - (q - b) = q^2 - 2qb + b^2 - q + b = (q^2 - q) + (b^2 + b - 2qb). \quad (7)$$

Comparing formulas (7) and (6), we note that with a decrease in the number of data vectors in the check group by the b value, the number of undetectable errors occurring within the check group under consideration has changed by the $(b^2 + b - 2qb) = b(b + 1 - 2q)$ value. The $b \in \{1, 2, \dots, 3^{m-k} - 1\}$ value, and the $q = 3^{m-k}$ value. The $b(b + 1 - 2q)$ expression for the marked b and q values are always less than zero. For example, $b = 3^{m-k} - 1$ (maximum value). Then we have:

$$(3^{m-k} - 1)((3^{m-k} - 1) + 1 - 2 \cdot 3^{m-k}) = (3^{m-k} - 1)(-3^{m-k}) < 0.$$

The number of undetectable errors in the check group has decreased by the $|b(b + 1 - 2q)|$ value.

In other check groups, the number of undetectable errors increases, since the total number of data vectors has increased by 1. Then in the groups with an increased number of vectors, the number of undetectable errors is:

$$(q + 1)(q + 1 - 1) = (q + 1)^2 - (q + 1) = q^2 + 2q + 1 - q - 1 = q^2 + q. \quad (8)$$

Comparing (8) and (6), we note that adding one vector to the check group increased the number of undetectable errors by the $2q$ value.

Since the number of groups with an increased number of data bits per unit is equal to b , the total increase in the number of undetectable errors is $2qb$. In the remaining check groups (in which there are q data vectors left) the number of undetectable errors persisted.

It follows from the reasoning that the number of undetectable errors has decreased due to a decrease in the number of vectors in one check group by $|b(b + 1 - 2q)|$ and an increase in the number of undetectable errors when vectors are added to other check groups by $2q$. The number of undetected errors has increased by $2q + b(b + 1 - 2q) = b^2 + b$. For $b = 1$, the number of undetectable errors increases by 2, for $b = 2$ – by 6, for $b = 3$ – by 12, etc. Adding two or number of data vectors to one check group leads to an even more significant increase in the number of undetectable errors.

Thus, even a minimal violation of the uniform distribution for all 3^m data vectors between all 3^k check vectors leads to an increase in the number of undetectable error codes. Hence, with the uneven-dimensional distribution of data vectors between the check vectors, it is impossible to reduce the number of un-

detectable errors, and the suggested assumption that code with a minimum total number of undetectable errors may have an uneven distribution incorrectly.

The theorem is proved.

In Tabl. 4, all data vectors $\langle f_1 f_2 f_3 \rangle$ $P(3,1)$ -code are distributed among all single-trit check vectors. The distribution is uniform.

TABLE IV. THE DISTRIBUTION OF DATA VECTORS TO CHECK GROUPS FOR $P(3,1)$ -CODE

g		
i	0	1
$\langle f_1 f_2 f_3 \rangle$		
$ii1$	iii	$ii0$
$i00$	$i01$	$i0i$
$i1i$	$i10$	$i11$
$0i0$	$0i1$	$0ii$
$00i$	000	001
011	$01i$	010
$1ii$	$1i0$	$1i1$
101	$10i$	100
110	111	$11i$

The check groups analysis corresponding to the check code vectors, allows you to calculate the total number of the undetectable error code. Each check group of such a code contains $\frac{3^m}{3} = 3^{m-1}$ data vectors. The distortion will not be detected by the code if the data vector of one group passes as a result of the distortion into the data vector of another group. Such transitions number within a group is $3^{m-1}(3^{m-1} - 1)$. Since there are three groups, the total number of undetectable errors in the considered code is $N_{m,1}^3 = 3 \cdot 3^{m-1}(3^{m-1} - 1) = 3^m(3^{m-1} - 1)$. The ternary parity code does not detect $N_{3,1}^3 = 3^3(3^{3-1} - 1) = 27 \cdot 8 = 216$ errors in the data vectors (among them 162 double and 54 triple errors). For comparison, the binary parity code does not detect $N_{m,1}^2 = 2^m(2^{m-1} - 1)$ errors in the data vectors [21]. For a three-bit vector, this number is equal to $N_{3,1}^2 = 24$ error, which is 9 times less than in the ternary parity code for the same length of the data vector.

Determine what errors proportion in the data vectors can be detected using the ternary parity code:

$$\lim_{m \rightarrow \infty} \gamma_{m,k} = \lim_{m \rightarrow \infty} \frac{N_{m,1}^3}{N_m^3} = \lim_{m \rightarrow \infty} \frac{3^m(3^{m-1} - 1)}{3^m(3^m - 1)} = \lim_{m \rightarrow \infty} \frac{3^{m-1} - 1}{3^m - 1} = \frac{1}{3} \quad (9)$$

The binary parity code does not detect 50% errors in the limit [21], whereas the ternary parity code – 33.3%.

It should be noted that the number of errors not detected by the ternary parity code is significant. However, unlike the binary parity code, for which no error of even multiplicity will be detected, for the ternary parity codes, the undetectable errors characteristics are much more diverse. For example, table 4 demonstrates that not every double error will lead to the check function value preservation.

Ternary parity codes can be used in the technical tools development for diagnosing digital devices operating in the ternary logic. At the same time, it is necessary to establish the errors features arising in the ternary data vectors, just as it was done in [29] for binary separable codes. At the same time, ternary parity codes detect any single errors. These codes can be applied, for example, when organizing check of logic circuits by groups of independent outputs (I-groups) [30]. Given the errors characteristics that occur at the outputs of the device (their multiplicities and types, determined by the number of values distortions combinations of i , 0 and 1, into false values of i^* , 0^* and 1^*), it is possible to simplify check devices and extend I-groups. However, this trend in the digital device synthesis theory in the ternary logic has not yet been investigated.

To design the $P(3,1)$ -code, the $g = f_1 \oplus f_2 \oplus f_3$ function is used. During the synthesis of the encoder, cascade connection of two modulo-three adders is required (in Fig. 1). To improve the error detection performance in the data vectors without introducing additional complexity into the encoder, the selection of two check trites can be used: $g_1 = f_1 \oplus f_2$ and $g_2 = f_2 \oplus f_3$. For the constructed $P(3,2)$ -code, the data vectors distribution between all check vectors is presented in Tabl. 5. It is also uniform, as for the $P(3,1)$ -code (this directly follows from the Theorem 1 formulation). According to the formula (3), the $P(3,2)$ -code will not detect 54 errors, which is 4 times less than the $P(3,1)$ -code. In addition, unlike the $P(3,1)$ -code, the $P(3,2)$ -code all vectors inside the check groups differ from each other in three bits (that is, they have Hamming code distance $d_{min}=3$), which means they do not detect only three-fold errors in the data vectors (a total of 54 three-fold errors). Thus, the distribution of undetectable errors by multiplicity, which is important, is biased towards errors of greater multiplicity.

TABLE V. THE DISTRIBUTION OF DATA VECTORS TO CHECK GROUPS FOR $P(3,2)$ -CODE

$\langle g_1 g_2 \rangle$								
ii	$i0$	$i1$	$0i$	00	01	$1i$	10	11
$\langle f_1 f_2 f_3 \rangle$								
$i0i$	$i00$	101	$i11$	$i1i$	$i10$	$ii0$	$ii1$	iii
$0i0$	$0i1$	$0ii$	$00i$	000	001	011	$01i$	010
111	$11i$	110	$1i0$	$1i1$	$1ii$	$10i$	100	101

Note that selecting two check trit method can be used similarly to the two groups check of device outputs. For codes with a large number of data bits, you can build modified parity codes with three or more check bits (including analogues of binary polynomial codes and Reed-Muller codes).

III. CONCLUSION

The study of the error detection features by ternary redundant codes focused on error detection is extremely important in the interested view in using devices build on ternary logic. The ternary codes using is promising when building devices with fault detection, in solving problems of offline and online diagnostics.

It should be noted that, as for binary logic devices, for ternary logic devices, the parity code is the simplest and allows to solve problems of fault detection with minimal hardware costs. Ternary parity codes refer to ternary codes with the lowest total number of undetectable errors in the data vectors for specific parameters m and k . However, as shown in the article, the

number of undetectable errors is large and amounts to 33.3% in the general case. For real digital devices, this indicator will be different, since the all possible values implementation at the outputs of devices, as a rule, is excluded. In this case, the parity code can be much more efficient.

In general, when errors are detected in ternary logic devices, the parity codes are more efficient than their binary counterparts. To increase detecting ability, several parity check functions can be used.

REFERENCES

- [1] W. Aspray "John von Neumann and the Origins of Modern Computing (History of Computing)", Boston, Cambridge: MIT, 1990, 396 p.
- [2] N.P. Brusencov, S.P. Maslov, V.P. Rozin, and A.M. Tishulina "Small Digital Computing Machine Setun" (in Russ), Moscow: Pub. House MGU, 1962, 140 p.
- [3] J. Connely "Ternary Computing Testbed 3-Trit Computer Architecture", California Polytechnic State University of San Luis Obispo, August 29th, 2008, 184 p.
- [4] D. Roy, and Jr. Merrill "Ternary Logic in Digital Computers", Proceedings of the SHARE design automation project (DAC '65), ACM New York, NY, USA, pp. 6.1-6.17, doi: 10.1145/800266.810759.
- [5] M. Hu, and K.C. Smith "Self-Checking Binary Logic Systems Using Ternary Logic Circuits", Canadian Electrical Engineering Journal, 1984, Vol. 9, Issue 3, Pp. 100-104, DOI: 10.1109/CEEJ.1984.6593793.
- [6] J. Wu "Ternary Logic Circuit for Error Detection and Error Correction", Proceedings of 19th International Symposium on Multiple-Valued Logic, 29-31 May 1989, Guangzhou, China, pp. 94-99, doi: 10.1109/ISMVL.1989.37766.
- [7] B.P. Lanyon, M. Barbieri, M.P. Almeida, T. Jennewein, T.C. Ralph, K.J. Resch, G.J. Pryde, J.L. O'Brien, A. Gilchrist and A.G. White "Simplifying Quantum Logic Using Higher-Dimensional Hilbert Spaces", Nature Physics, 2009, Vol. 5, Issue 2, pp. 134-140, doi: 10.1038/nphys1150.
- [8] B. Cambou, P.G. Flikkema, J. Palmer, D. Telesca, and C. Philabaum "Can Ternary Computing Improve Information Assurance?", Cryptography, 2018, Volume 2, Issue 1 (March 2018), pp. 1-16, doi: 10.3390/cryptography2010006.
- [9] C. Vudadha, S. Katragadda, and P.S. Phaneendra "2:1 Multiplexer Based Design for Ternary Logic Circuits", IEEE Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia), 19-21 December 2013, Visakhapatnam, India, pp. 46-51, doi: 10.1109/PrimeAsia.2013.6731176.
- [10] S. Ahmad, and M. Alam "Balanced-Ternary Logic for Improved and Advanced Computing", International Journal of Computer Science and Information Technologies (IJCSIT), 2014, Vol. 5, Issue 4, pp. 5157-5160.
- [11] R.S.P. Nair, S.C. Smith, and J. Di "Delay-Insensitive Ternary CMOS Logic for Secure Hardware", Journal of Low Power Electronics and Applications, 2015, Issue 5, pp. 183-215, doi:10.3390/jlpea5030183.
- [12] R.N. Uma Mahesh, and J. Sudeep "Design and Novel Approach for Ternary and Quaternary Logic Circuits", 2nd International Conference for Convergence in Technology (I2CT), 7-9 April 2017, Mumbai, India, pp. 1224-1227, doi: 10.1109/I2CT.2017.8226322.
- [13] S. Kim, T. Lim, and S. Kang "An Optimal Gate Design for the Synthesis of Ternary Logic Circuits", 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), 22-25 January 2018, Jeju, South Korea, pp. 476-481, doi: 10.1109/ASPAC.2018.8297369.
- [14] C. Vudadha, S. Rajagopalan, A. Dusi, P.S. Phaneendra, and M.B. Srinivas "Encoder-Based Optimization of CNFET-Based Ternary Logic Circuits", IEEE Transactions on Nanotechnology, 2018, Vol. 17, Issue 2, Pp. 299-310, DOI: 10.1109/TNANO.2018.2800015.
- [15] Md.R. Rahman, and J.E. Rise "On Designing a Ternary Reversible Circuit for Online Testability", IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 2011, pp. 1-7, doi: 10.1109/PACRIM.2011.6032878.
- [16] N.M. Nayeem, and J.E. Rice "Design of an Online Testable Ternary Circuit from the Truth Table", Lecture Notes in Computer Science book series (LNCS, volume 7581): Reversible Computation, 4th International Workshop on Reversible Computation (RC 2012), Copenhagen, Denmark, July 2-3, 2012, pp. 152-159.
- [17] S.J. Piestrak "Design of Self-Testing Checkers for Unidirectional Error Detecting Codes", Wrocław: Ofiyna Wydawnicza Politechniki Wrocławskiej, 1995, 111 p.
- [18] M. Goessel, V. Ocheretny, E. Sogomonyan, and D. Marienfeld "New Methods of Concurrent Checking: Edition 1", Dordrecht: Springer Science+Business Media B.V., 2008, 184 p.
- [19] D. Efanov, V. Sapozhnikov, and Vl. Sapozhnikov "Generalized Algorithm of Building Summation Codes for the Tasks of Technical Diagnostics of Discrete Systems", Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, September 29 - October 2, 2017, pp. 365-371, doi: 10.1109/EWDTS.2017.8110126.
- [20] R.F. Mirzaee, M.S. Daliri, K. Navi, and N. Bagherzadeh "A Single Parity-Check Digit for One Trit Error Detection in Ternary Communication Systems: Gate-Level and Transistor-Level Designs", Journal of Multiple-Valued Logic and Soft Computing, 2017, 29 (3-4), pp. 303-326.
- [21] V. Sapozhnikov, Vl. Sapozhnikov, and D. Efanov "Modular Sum Code in Building Testable Discrete Systems", Proceedings of 13th IEEE East-West Design & Test Symposium (EWDTS'2015), Batumi, Georgia, September 26-29, 2015, pp. 181-187, doi: 10.1109/EWDTS.2015.7493133.
- [22] A.E. Brouwer, H.O. Hamalainen, P.R.J. Ostergard, and N.J.A. Sloane "Bounds on Mixed Binary/Ternary Codes", IEEE Transactions on Information Theory, 1988, vol. 44, Issue 1 pp. 140-161, doi: 10.1109/18.651001.
- [23] T.A. Gulliver, and P.R.J. Ostergard "Improved Bounds for Ternary Linear Codes of Dimension 7", IEEE Transactions on Information Theory, 1997, Vol 43, Issue 4, pp. 1377-1381, doi: 10.1109/18.605613.
- [24] N. Bitouze, A. Graell i Amat, and E. Rosnes "Error-Correcting Coding for a Nonsymmetric Ternary Channel", IEEE Transactions on Information Theory, 2010, Vol. 56, Issue 11, pp. 5715-5729, doi: 10.1109/TIT.2010.2069211.
- [25] A. Laaksonen, and P.R.J. Östergård "New Lower Bounds on Error-Correcting Ternary, Quaternary and Quinary Codes", Lecture Notes in Computer Science 10495, Springer: Coding Theory and Applications. 5th International Castle Meeting, ICMCTA 2017, Vihula, Estonia, August 28-31, 2017, pp. 228-237
- [26] M. Svanström "A Lower Bound for Ternary Constant Weight Codes", IEEE Transactions on Information Theory, 1997, vol. 43, pp. 1630-1632.
- [27] M. Svanström, P.R.J. Östergård, and G.T. Bogdanova "Bounds and Constructions for Ternary Constant-Composition Codes", IEEE Transactions on Information Theory, 2002, vol. 48, pp. 101-111.
- [28] D.A. Pospelov "Logical Methods of Analysis and Synthesis of Circuits" (in Russ), Moscow: Energy, 1974, 368 p.
- [29] V.V. Sapozhnikov, Vl.V. Sapozhnikov, and D.V. Efanov "Errors Classification in Information Vectors of Systematic Codes" (in Russ.), Journal of Instrument Engineering, 2015, Vol. 58, Issue 5, pp. 333-343. DOI 10.17586/0021-3454-2015-58-5-333-343.
- [30] D. Efanov, V. Sapozhnikov, and Vl. Sapozhnikov "Synthesis of Self-Checking Combinational Devices Based on Allocating Special Groups of Outputs", Automation and Remote Control, 2018, issue 9, pp. 1607-1618, doi: 10.1134/S0005117918090060.

Fast and Secure Unified Field Multiplier for ECC Based on the 4-Segment Karatsuba Multiplication

Ievgen Kabin, Zoya Dyka, Dan Klann and Peter Langendoerfer
IHP – Leibniz-Institut für innovative Mikroelektronik
Im Technologiepark 25
Frankfurt (Oder), Germany

Abstract— In this paper we introduce an accelerator for the ECC kP operation in two different types of Galois fields i.e. $GF(p)$ and $GF(2^n)$. In order to ensure fast execution of the multiplication in both cases we integrated the carry bit separation technique to speed-up the multiplication in $GF(p)$. The two most important contributions of this paper are that the partial multiplier applied is used for both types of Galois field and the second one is that our design is resistant against Horizontal Collision Correlation Analysis. The latter was verified in 20 test runs per supported elliptic curve.

Keywords— ECC; Unified Field Multiplier; side channel analysis (SCA) attack; horizontal attacks.

I. INTRODUCTION

The Internet of Things (IoT) is considered to be the next global trend that will revolutionize our lives. The number of devices is predicted to reach more than 20 billion in the next few years. The sheer amount of devices clearly determines that the vast majority of these devices cannot be operated by human beings but will be embedded devices hidden in our environment, controlling e.g. medical devices. In addition it means that we all will rely on the proper operation of these devices. The point here is that no one can guarantee any system properties if the basic security features such as confidentiality, integrity and availability cannot be ensured. In this paper we focus on the integrity feature as this essential for guaranteeing proper functionality in the sense of taking decisions based on correct sensed data e.g. temperature etc. In addition integrity allows to verify which device sent the data i.e. authenticity can be proven.

As we are taking about embedded devices the efficiency of the solutions used to ensure security features is of utmost importance. This is the reason why we decided to research implementations of Elliptic Curve Cryptography (ECC) as this system is considered to be by far more efficient than e.g. RSA.

The elliptic curve digital signature algorithm (ECDSA) [1] can be used for ensuring integrity. To do so each of the 20 billion devices needs to use its ECC private key. As the devices may be deployed in the wild there is a significantly high probability that the devices can be attacked using side channel

attacks that exploit measurements of physical parameters to extract the private keys. As according to Kerckhoff's principle the private key has to be kept secret, such an attack poses a significant threat to the system, i.e. misusing the key can cause dangerous situations in the traffic, car crashes, etc.

In order to avoid successful malicious attacks the cryptographic implementations need to be protected against side channel analysis (SCA)¹ attacks. This is a highly demanding task, especially if the design shall support the two different kinds of ECC i.e. elliptic curves over $GF(2^m)$ and $GF(p)$.

Hardware implementations for performing multiplications in both types of fields using the same data path (i.e. a unified architecture) were discussed in literature. In most of the cases they are designed and analyzed from the time and area efficiency and complexity points of view. However, their resistance against SCA attacks are still not investigated so far.

In this paper we describe our implementation of a field multiplier that can calculate the product for 4 NIST elliptic curves, over $GF(p)$ as well as over $GF(2^n)$ and we discuss its resistance against the Horizontal Collision Correlation Analysis (HCCA) attack [4]. The two most important contributions of this paper are that the partial multiplier (PM) applied is used for both types of Galois field and the second one is that our design is resistant against HCCA. The rest of this paper is structured as follows. In section II we give implementation details of our unified multipliers. We discuss the performed horizontal attacks in section III. The paper finishes with short conclusions.

II. IMPLEMENTED FIELD MULTIPLIERS

The multiplication is the most complex field operation in ECC hardware accelerators and in our design too. The polynomial multiplication in $GF(2^n)$ and the multiplication of two large binary integer numbers in $GF(p)$ (i.e. the first step of the field multiplications in Galois fields) can be realized by applying the classical multiplication method.

We implemented our field multiplier using the 4-segment Karatsuba multiplication method [5], [6]. Both multiplicands A and B are segmented into four parts: A_3, A_2, A_1, A_0 and B_3, B_2, B_1, B_0 respectively. Due to the fact that the multiplier can

¹ Please see e.g. [2] or [3] for a very condensed overview of different types of SCAs and their classification.

calculate the product of up to 283 bit long operands the parts A_2, A_1, A_0 and B_2, B_1, B_0 are 71 bit long always. The parts A_3 and B_3 in our hardware implementation are up to 70 bit long. It is applicable for the multiplication of 283 bit operands as well as for operands of a smaller length. For example if a product of 283 bit long operands is calculated the A_3 and B_3 are up to 70 bit long. If a product of 224 bit long operands has to be calculated the segments of the multiplicands A_3 and B_3 are only 11 bit long². Formula (1) is the 4-segment Karatsuba Multiplication Method (MM) for two binary polynomials $A(t)$ and $B(t)$ and formula (2) is the 4-segment Karatsuba MM for two large binary integer numbers A and B .

$$\begin{aligned}
A(t) \cdot B(t) &= A_3 A_2 A_1 A_0 \cdot B_3 B_2 B_1 B_0 = \\
&= (A_3 \cdot 2^{3m} \oplus A_2 \cdot 2^{2m} \oplus A_1 \cdot 2^{1m} \oplus A_0) \cdot (B_3 \cdot 2^{3m} \oplus B_2 \cdot 2^{2m} \oplus B_1 \cdot 2^{1m} \oplus B_0) = \\
&= S_6 \cdot 2^{6m} \oplus S_5 \cdot 2^{5m} \oplus S_4 \cdot 2^{4m} \oplus S_3 \cdot 2^{3m} \oplus S_2 \cdot 2^{2m} \oplus S_1 \cdot 2^{1m} \oplus S_0 = \\
&= A_0 B_0 \cdot 2^0 \oplus (A_0 B_0 \oplus A_1 B_1 \oplus (A_0 \oplus A_1)(B_0 \oplus B_1)) \cdot 2^{1m} \\
&\oplus (A_0 B_0 \oplus A_1 B_1 \oplus A_2 B_2 \oplus (A_0 \oplus A_2)(B_0 \oplus B_2)) \cdot 2^{2m} \\
&\oplus \left(\begin{array}{l} A_0 B_0 \oplus A_1 B_1 \oplus A_2 B_2 \oplus A_3 B_3 \oplus \\ (A_0 \oplus A_2)(B_0 \oplus B_2) \oplus (A_0 \oplus A_1)(B_0 \oplus B_1) \oplus \\ (A_1 \oplus A_3)(B_1 \oplus B_3) \oplus (A_2 \oplus A_3)(B_2 \oplus B_3) \oplus \\ (A_0 \oplus A_1 \oplus A_2 \oplus A_3)(B_0 \oplus B_1 \oplus B_2 \oplus B_3) \end{array} \right) \cdot 2^{3m} \\
&\oplus (A_1 B_1 \oplus A_2 B_2 \oplus A_3 B_3 \oplus (A_1 \oplus A_3)(B_1 \oplus B_3)) \cdot 2^{4m} \\
&\oplus (A_2 B_2 \oplus A_3 B_3 \oplus (A_2 \oplus A_3)(B_2 \oplus B_3)) \cdot 2^{5m} \oplus A_3 B_3 \cdot 2^{6m} = \\
&= p_1 \cdot 2^0 \oplus (p_1 \oplus p_2 \oplus p_7) \cdot 2^{1m} \oplus (p_1 \oplus p_2 \oplus p_3 \oplus p_5) \cdot 2^{2m} \oplus \\
&\oplus (p_1 \oplus p_2 \oplus p_3 \oplus p_4 \oplus p_5 \oplus p_7 \oplus p_6 \oplus p_8 \oplus p_9) \cdot 2^{3m} \oplus \\
&\oplus (p_2 \oplus p_3 \oplus p_4 \oplus p_6) \cdot 2^{4m} \oplus (p_3 \oplus p_4 \oplus p_8) \cdot 2^{5m} \oplus p_4 \cdot 2^{6m} = \\
&= C_7 C_6 C_5 C_4 C_3 C_2 C_1 C_0,
\end{aligned}$$

with partial products:

$$\begin{aligned}
p_1 &= A_0 B_0, \quad p_2 = A_1 B_1, \quad p_3 = A_2 B_2, \quad p_4 = A_3 B_3, \\
p_5 &= (A_0 \oplus A_2)(B_0 \oplus B_2), \quad p_6 = (A_1 \oplus A_3)(B_1 \oplus B_3), \\
p_7 &= (A_0 \oplus A_1)(B_0 \oplus B_1), \quad p_8 = (A_2 \oplus A_3)(B_2 \oplus B_3), \\
p_9 &= (A_0 \oplus A_1 \oplus A_2 \oplus A_3)(B_0 \oplus B_1 \oplus B_2 \oplus B_3)
\end{aligned}$$

each partial product p_i is $2m-1$ bit long,
each sum S_i is $2m-1$ bit long,
each segment of the product C_6, \dots, C_0 is m bit long,
the segment C_7 is $m-1$ bit long.

Formulae (1) and (2) are similar. Formula (1) can be obtained from (2) if the bitwise XOR operation will be applied instead the addition and subtraction in formula (2). The multiplication formulae (1) and (2) contain 9 partial products each. In each clock cycle only one of 9 partial products is calculated. The multiplier takes 9 clock cycles to calculate the field product.

TABLE I. represents formula (2) and shows the sequence of additions/subtractions of m bit long segments required for the accumulation of partial products p_j . Each partial product p_j has to be subtracted ($p_{j(-)}$) or added ($p_{j(+)}$) to the result. Partial products p_5 to p_8 marked as grey and p_9 marked as black in the left column are $2m+2$ and $2m+4$ bits long respectively. The calculation sequence for the partial products is shown in the most right column by indicating the clock cycle $clk^1 \dots clk^9$ in

which corresponding additions and/or subtractions are performed.

$$\begin{aligned}
A \cdot B &= A_3 A_2 A_1 A_0 \cdot B_3 B_2 B_1 B_0 = \\
&= (A_3 \cdot 2^{3m} + A_2 \cdot 2^{2m} + A_1 \cdot 2^{1m} + A_0) \cdot (B_3 \cdot 2^{3m} + B_2 \cdot 2^{2m} + B_1 \cdot 2^{1m} + B_0) = \\
&= s_6 \cdot 2^{6m} + s_5 \cdot 2^{5m} + s_4 \cdot 2^{4m} + s_3 \cdot 2^{3m} + s_2 \cdot 2^{2m} + s_1 \cdot 2^{1m} + s_0 = \\
&\left(\begin{array}{l} p_1 \cdot 2^0 + p_7 \cdot 2^{1m} + (p_2 + p_5) \cdot 2^{2m} + \\ + (p_1 + p_2 + p_3 + p_4 + p_9) \cdot 2^{3m} + \\ + (p_3 + p_6) \cdot 2^{4m} + p_8 \cdot 2^{5m} + p_4 \cdot 2^{6m} \end{array} \right) - \left(\begin{array}{l} (p_1 + p_2) \cdot 2^{1m} + (p_1 + p_3) \cdot 2^{2m} + \\ (p_5 + p_7 + p_6 + p_8) \cdot 2^{3m} + \\ + (p_2 + p_4) \cdot 2^{4m} + (p_3 + p_4) \cdot 2^{5m} \end{array} \right) = \quad (2) \\
&= C_7 C_6 C_5 C_4 C_3 C_2 C_1 C_0,
\end{aligned}$$

with partial products:

$$\begin{aligned}
p_1 &= A_0 B_0, \quad p_2 = A_1 B_1, \quad p_3 = A_2 B_2, \quad p_4 = A_3 B_3, \quad \left. \begin{array}{l} \text{each } 2m \text{ bits long} \\ p_5 = (A_0 + A_2)(B_0 + B_2), \quad p_6 = (A_1 + A_3)(B_1 + B_3), \\ p_7 = (A_0 + A_1)(B_0 + B_1), \quad p_8 = (A_2 + A_3)(B_2 + B_3), \end{array} \right\} \text{each } 2(m+1) \text{ bits long} \\
p_9 &= (A_0 + A_1 + A_2 + A_3)(B_0 + B_1 + B_2 + B_3) \quad \left. \begin{array}{l} \text{each segment of the product } C_i \text{ is } m \text{ bits long.} \end{array} \right\} 2(m+2) \text{ bits long}
\end{aligned}$$

TABLE I. REPRESENTATION OF FORMULA (2).

	C_7	C_6	C_5	C_4	C_3	C_2	C_1	C_0		
p_1				+					$p_{1(+)}$	clk^5
					-				$p_{1(-)}$	
p_2					+				$p_{2(+)}$	clk^6
						-			$p_{2(-)}$	
p_3						+			$p_{3(+)}$	clk^9
							-		$p_{3(-)}$	
p_4							+		$p_{4(+)}$	clk^1
								-	$p_{4(-)}$	
p_5								+	$p_{5(+)}$	clk^8
									$p_{5(-)}$	
p_6									$p_{6(+)}$	clk^3
									$p_{6(-)}$	
p_7									$p_{7(+)}$	clk^7
									$p_{7(-)}$	
p_8									$p_{8(+)}$	clk^2
									$p_{8(-)}$	
p_9									$p_{9(+)}$	clk^4

As for $GF(2^n)$ all subtractions and additions are bitwise XOR operations we adapted the formula of the 4-segment Karatsuba MM for the calculation of $GF(p)$ products by applying the Carry Bit Separation (CBS) technique. The structure of the implemented field multiplier is shown on Fig. 1. Easy examples shown in Fig. 2 and in Fig. 3 explain the CBS techniques for a multiplier and an adder.

Each sum s_i in Fig. 2 has its own length, for example s_0 is 1 bit long, s_1 is 2 bits long, s_3 is 3 bits long: $s_0 = s_0^0$; $s_1 = s_1^1 s_1^0$; $s_3 = s_3^2 s_3^1 s_3^0$. A calculation of all carry values can be performed separately. Therefore a sum in $GF(p)$ can be represented as a sum of the sum in $GF(2^n)$ and the carry values. The partial product calculation using the CBS technique can be described as: $p = p^{\text{XOR}} + sel \cdot p^{\text{carry}}$.

² 224-3*71=11

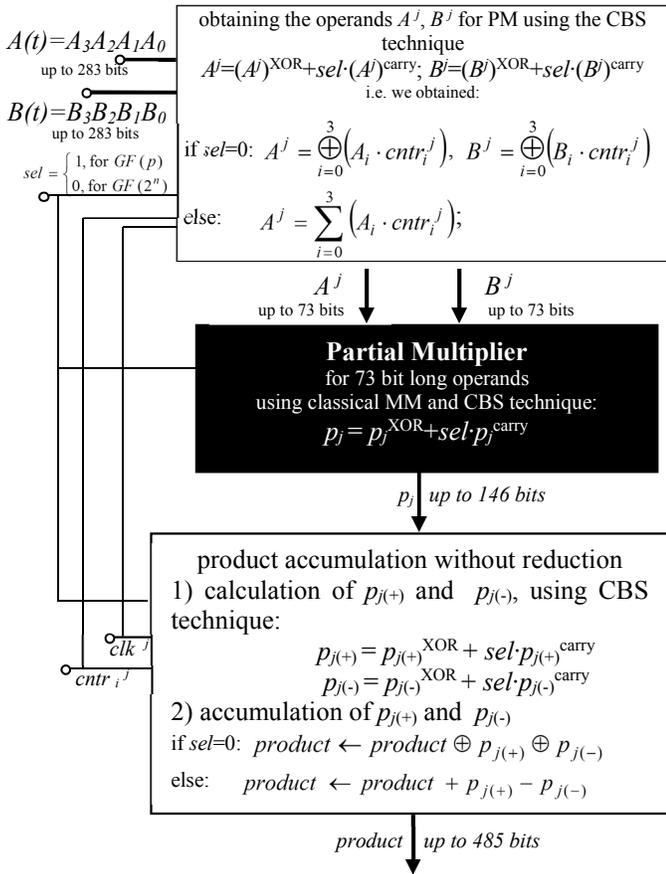


Fig. 1. Unified field multiplier without reduction implemented by applying the 4-segment Karatsuba MM with the Carry Bit separation technique.

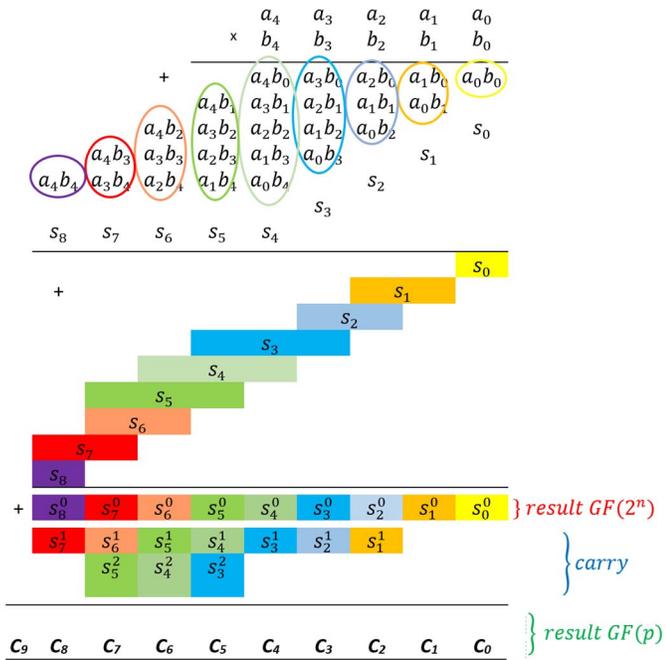


Fig. 2. Multiplication using the CBS technique on the example of 5-bit long multiplicands, using the classical MM.

$$\begin{array}{r}
 + \quad a_2 \quad a_1 \quad a_0 \\
 \quad b_2 \quad b_1 \quad b_0 \\
 \hline
 c_2^{XOR} \quad c_1^{XOR} \quad c_0^{XOR} \quad \} \text{result } GF(2^n): C^{XOR} \\
 + \quad c_2^{carry} \quad c_1^{carry} \quad c_0^{carry} \quad 0 \quad \} \text{carry: } C^{carry} \\
 \hline
 c_3 \quad c_2 \quad c_1 \quad c_0 \quad \} \text{result } GF(p): \\
 C = C^{XOR} + C^{carry}
 \end{array}$$

Fig. 3. Addition using the CBS technique on the example of a 3-bit adder: $C^{XOR} = c_2^{XOR} c_1^{XOR} c_0^{XOR}$, where $c_i^{XOR} = a_i \oplus b_i$; $C^{carry} = c_2^{carry} c_1^{carry} c_0^{carry} 0$, where $c_i^{carry} = a_i \cdot b_i$, for $0 \leq i \leq 2$. The addition can be described as follows: $C = C^{XOR} + sel \cdot C^{carry}$, with $sel=1$ if a sum in $GF(p)$ has to be calculated and $sel=0$ for $GF(2^n)$.

A. Optimizing the Multiplier

We synthesized the multiplier described above for the IHP 250 nm gate technology [7] for a frequency of 10 MHz only. It was the maximum frequency for which the synthesis was completed successfully [3]. Due to the fact that the synthesized multiplier was slow without any reduction, we made different optimizations of the partial multiplier with the goal to reduce the time for the signal propagation. Applying different techniques described in [8] increased the frequency significantly. In one of our experiments we implemented two partial multipliers as parts of the serial 9 clock cycles multiplier: the first one calculated the partial product in $GF(p)$ and the second one in $GF(2^n)$. It was the multiplier with the best parameters (area and frequency) from all our experiments. We don't have any knowledge about the optimization algorithms applied by the Synopsys synthesis tools but we selected this multiplier for our further experiments.

B. Implementing Reductions

As our unified multiplier requires modulo reduction in prime and binary extension fields, we implemented 4 reductions. The reductions for both binary curves were implemented as packages. The reduction for the ECs P-224 and P-256 are implemented as a single block each. All reductions are parts of the field multiplier. The reduction is performed in each clock cycle after the calculation of the new partial product.

1) Reduction for B-Curves

The reduction is an essential operation in the calculation of the field product in $GF(2^n)$. The polynomial product in $GF(2^{233})$ is a polynomial of degree 464 that can be represented as a 465 bit long number:

$$C'(t) = \sum_{i=0}^{464} c_i t^i = c_{464} c_{463} \dots c_2 c_1 c_0.$$

The irreducible polynomial in $GF(2^{233})$ is: $f(t) = t^{233} + t^{74} + 1$.

The result of the reduction is a polynomial $P(t)$ that can be represented as a 233 bit long number:

$$P(t) = C'(t) \bmod f(t) = \sum_{i=0}^{232} p_i t^i = p_{232} p_{231} \dots p_2 p_1 p_0$$

The reduction of each monomial of a degree from 233 up to 465 results in:

$$\begin{aligned}
 t^{233} &= t^{74} + 1 \\
 t^{234} &= t^{75} + t \\
 t^{235} &= t^{76} + t^2
 \end{aligned}$$

$$\begin{aligned}
&\dots \\
t^{391} &= t^{232} + t^{158} \\
t^{392} &= t^{233} + t^{159} = t^{74} + I + t^{159} = t^{159} + t^{74} + I \\
t^{393} &= t^{160} + t^{75} + t \\
t^{394} &= t^{161} + t^{76} + t^2 \\
&\dots \\
t^{464} &= t^{231} + t^{146} + t^{72}
\end{aligned}$$

Thus, the reduction requires 318+219=537 XOR gates only, whereby it is a fast operation. The number of the gates can be further reduced (and is made automatically by the Synopsys synthesis tools) exploiting the fact that some of the XOR gates have the same operands. To illustrate this the reduction can be schematically represented as a table (see TABLE II.):

TABLE II. REPRESENTATION OF THE REDUCTION AS A TABLE.

bitwise XOR	$C_{232} C_{231} \dots C_{159} C_{158} C_{157} \dots C_{147} C_{146} C_{145} \dots C_{77} C_{76} C_{75} C_{74} C_{73} C_{72} \dots C_1 C_0$
	$C_{391} C_{390} \dots C_{380} C_{379} C_{378} \dots C_{310} C_{309} C_{308} C_{307} C_{306} C_{305} \dots C_{234} C_{233}$
	$C_{391} C_{390} \dots C_{318} C_{317} C_{316} \dots C_{306} C_{305} C_{304} \dots C_{236} C_{235} C_{234} C_{233}$
	$C_{464} \dots C_{392}$
Field product: $P(t) =$	$D_{232} D_{231} \dots D_{159} D_{158} D_{157} \dots D_{147} D_{146} D_{145} \dots D_{77} D_{76} D_{75} D_{74} D_{73} D_{72} \dots D_1 D_0$

The pairs of operands to be XORed marked with the same colour in TABLE II. are identical. Thus it is sufficient to calculate the result of the XOR operation for each pair only once. So, 73 XOR operations can be saved. This results in 537-73=464 XOR gates for realizing the complete reduction unit for the EC B-233.

We implemented the reduction for the EC B-283 in a similar way as for the EC B-233. The irreducible polynomial is the pentanomial $f(t) = t^{283} + t^{12} + t^7 + t^5 + I$ [1].

1) Reduction for P-Curves

We implemented the modulo reduction for prime curves according to the NIST standard [9]. The implemented algorithms for both elliptic curves used are shown below, see Algorithm 1 and Algorithm 2.

Algorithm 1: Modulo reduction algorithm for P-224

Input: integer $c = \{c_{13}, c_{12}, \dots, c_1, c_0\}$, $0 \leq c < p_{224}^2$, where $p_{224} = 2^{224} - 2^{96} + 1$, each c_i is 32-bit integer

$k_1 \dots k_5$ are the products of concatenation of 32-bit integers:

$$k_1 = \{c_6, c_5, c_4, c_3, c_2, c_1, c_0\}$$

$$k_2 = \{c_{10}, c_9, c_8, c_7, 0, 0, 0\}$$

$$k_3 = \{0, c_{13}, c_{12}, c_{11}, 0, 0, 0\}$$

$$k_4 = \{c_{13}, c_{12}, c_{11}, c_{10}, c_9, c_8, c_7\}$$

$$k_5 = \{0, 0, 0, 0, c_{13}, c_{12}, c_{11}\}$$

Output:

$$c \bmod p_{224} = (k_1 + k_2 + k_3 - k_4 - k_5) \bmod p_{224}$$

Algorithm 2: Modulo reduction algorithm for P-256

Input: integer $c = \{c_{15}, c_{12}, \dots, c_1, c_0\}$, $0 \leq c < p_{256}^2$, where $p_{256} = 2^{256} - 2^{224} + 2^{192} + 2^{96} - 1$, each c_i is 32-bit integer

$k_1 \dots k_9$ are the products of concatenation of 32-bit integers:

$$k_1 = \{c_7, c_6, c_5, c_4, c_3, c_2, c_1, c_0\}$$

$$k_2 = \{c_{15}, c_{14}, c_{13}, c_{12}, c_{11}, 0, 0, 0\}$$

$$k_3 = \{0, c_{15}, c_{14}, c_{13}, c_{12}, 0, 0, 0\}$$

$$k_4 = \{c_{15}, c_{14}, 0, 0, 0, c_{10}, c_9, c_8\}$$

$$k_5 = \{c_8, c_{13}, c_{15}, c_{14}, c_{13}, c_{11}, c_{10}, c_9\}$$

$$k_6 = \{c_{10}, c_8, 0, 0, 0, c_{13}, c_{12}, c_{11}\}$$

$$k_7 = \{c_{11}, c_9, 0, 0, c_{15}, c_{14}, c_{13}, c_{12}\}$$

$$k_8 = \{c_{12}, 0, c_{10}, c_9, c_8, c_{15}, c_{14}, c_{14}\}$$

$$k_9 = \{c_{13}, 0, c_{11}, c_{10}, c_9, 0, c_{15}, c_{14}\}$$

Output:

$$c \bmod p_{256} = (k_1 + 2k_2 + 2k_3 + k_4 + k_5 - k_6 - k_7 - k_8 - k_9) \bmod p_{256}$$

The last step in these algorithms requires an additional modulo reduction. For P-224 the expression $(k_1 + k_2 + k_3 - k_4 - k_5)$ has a maximum value in case, k_4 and k_5 are equal to zero and a minimum value when k_1, k_2 and k_3 are equal to zero. I.e. $(k_1 + k_2 + k_3 - k_4 - k_5) \in (3 \cdot p_{224}; -2 \cdot p_{224})$.

Therefore, to speed up the calculation of this last step we calculated five values simultaneously:

$$S_{-2} = (k_1 + k_2 + k_3 - k_4 - k_5) + 2 \cdot p_{224}$$

$$S_{-1} = (k_1 + k_2 + k_3 - k_4 - k_5) + p_{224}$$

$$S_0 = (k_1 + k_2 + k_3 - k_4 - k_5)$$

$$S_1 = (k_1 + k_2 + k_3 - k_4 - k_5) - p_{224}$$

$$S_2 = (k_1 + k_2 + k_3 - k_4 - k_5) - 2p_{224}$$

Each of the values $S_{-2}, S_{-1}, S_0, S_1, S_2$ can be longer than 224 bits. The first number in the sequence $S_{-2}, S_{-1}, S_0, S_1, S_2$ that is not longer than 224 bit is the reduced result, i.e. the output.

For the P-256 reduction algorithm $(k_1 + 2k_2 + 2k_3 + k_4 + k_5 - k_6 - k_7 - k_8 - k_9) \in (7 \cdot p_{256}; -4 \cdot p_{256})$ it is necessary to calculate simultaneously eleven values i.e. $S_{-4}, S_{-3}, S_{-2}, S_{-1}, S_0, S_1, S_2, S_3, S_4, S_5, S_6$. Each of these values can be longer than 256 bits. The first number in the sequence $S_{-4}, S_{-3}, S_{-2}, S_{-1}, S_0, S_1, S_2, S_3, S_4, S_5, S_6$ that is not longer than 256 bit is the reduced result, i.e. the output.

The modulo reduction operation for all four curves in our implementation is done within a single clock cycle.

C. Parameters of the synthesized Multipliers

We synthesized our optimized unified polynomial multipliers for the IHP 250 nm technology for two frequencies:

- for 10 MHz with the goal to compare it to the not optimized multiplier (without implemented reductions);
- for 71 MHz, that was the maximum frequency for that the multiplier was synthesized.

- Parameters of the synthesized unified field multiplier are shown in TABLE III, as well as information regarding original field multiplier before the optimizations.

TABLE III. PARAMETERS OF SYNTHESIZED MULTIPLIERS.

	our unified multiplier		
	before optimization	after optimization	after optimization
frequency	10 MHz	10 MHz	71 MHz
MM	4-segment Karatsuba		
area	3.2 mm ²	4.14 mm ²	5.41 mm ²
power	61.3mW (B-233) 84.2 mW (P-256)	64.7mW (B-233) 73.7mW (P-256)	78.1mW (B-233) 82.7mW (P-256)
Partial Mult.	73 bit long inputs		
reduction	area 1.785 mm ²	area 1.150 mm ²	area 1.210 mm ²
	no	yes	yes

It can be seen that the area of the PM after the optimization of the Synopsys Tools for 10 MHz was reduced significantly (about 36 per cent). Consequently, its signal delay path was reduced too. So, it was possible to synthesize it for a 14 ns clock cycle period that corresponds to 71 MHz clock frequency.

III. PERFORMED HORIZONTAL ATTACKS

Here we describe how we performed the Horizontal Collision Correlation Analysis attack [4] against our field multiplier.

A. Bauer attack

In 2013 A. Bauer et al [4] published their HCCA attack on elliptic curves over prime fields $GF(p)$. The attack is based on the assumption that two multiplications with one common multiplicand are distinguishable from two multiplications with different multiplicands. If the assumption is true. An EC point doubling can be distinguished from an EC point addition even in double-and-add kP algorithms implemented according to the atomicity principle [10]. Thus, this knowledge can be used for revealing the key. The field multiplier analysed in the attack in [4] was realized using the classical multiplication formula. Experimental results in [4] confirm the assumption about the distinguishability of such multiplications: Pearson's coefficients calculated for traces of two multiplications with a common multiplicand differ significantly from Pearson's coefficients calculated for two multiplications with completely different multiplicands.

This type of attack cannot reveal keys when binary curves are attacked, but it can help to separate the trace into parts which correspond to the processing of a single bit of the key. We denote such parts further as slots. The fact that a multiplication with the parameter b of the elliptic curve or a multiplication with the x coordinate of the input point P is executed in each slot can be exploited to segment the power trace into slots.

In order to investigate the resistance of our unified multiplier against HCCA [4] we run the following experiments:

- Simulation of power traces of the product calculation for 4 multiplications with one common and two completely different operands: $mult_1=a \cdot b$; $mult_2=c \cdot d$; $mult_3=a \cdot e$; $mult_4=f \cdot g$.

- Calculation of Pearson coefficients k_i ($1 \leq i \leq 4$) using the power shape of the following product calculations: k_1 using $mult_1$ and $mult_3$, here operand a is common in 2 multiplications; k_2 using $mult_2$ and $mult_4$, k_3 using $mult_1$ and $mult_2$ and k_4 using $mult_2$ and $mult_3$. Coefficients k_2 , k_3 and k_4 correspond to multiplications with different operands.

The difference of the coefficient k_1 to the coefficients of k_3 and/or k_4 , clearly indicates that the multiplications with a common operand $mult_1$ and $mult_3$ are distinguishable from product calculations with different operands such as $mult_1$ and $mult_2$ and/or $mult_1$ and $mult_4$. Please note that this distinguishability has to be observed for each experiment (see steps 1) and 2)) for a successful HCCA attack.

We performed the experiment described above for operands of length 224 bits for P-224, 233 bits for B-233, 256 bits for P-256 and for 283 bits for B-283 20 times each. The Synopsys PrimeTime tool [11] was used for the simulations. We generated 283 bit long random numbers. We used the least significant n bits of the 283 bit long numbers for operands of smaller length n . Fig. 4 shows coefficients k_1 , k_2 , k_3 and k_4 calculated for all experiments.

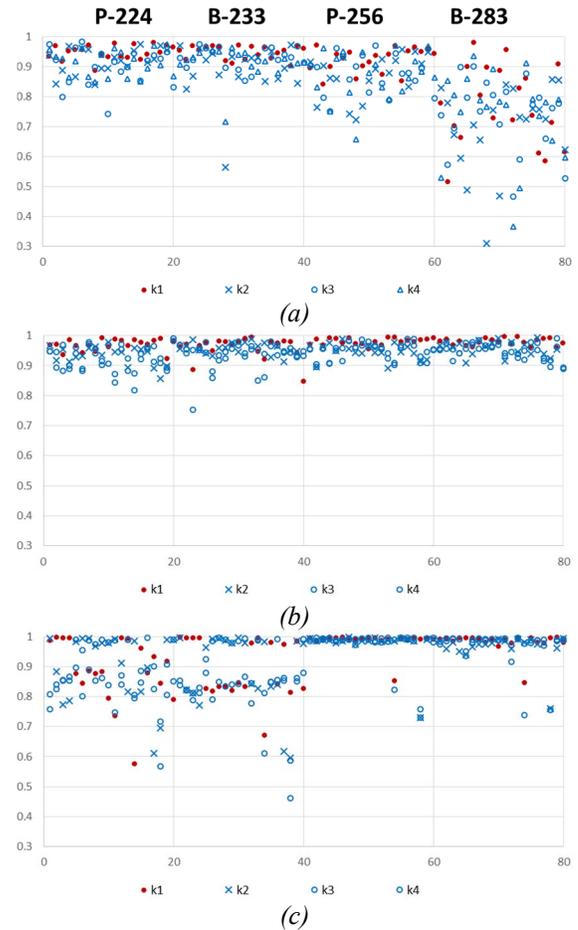


Fig. 4. Pearson coefficients: four coefficients per experiment are calculated; 20 experiments for each of the 4 investigated ECs are run. x axis shows the number of experiments counted continuously from 0 to 80: 0-20 corresponds to runs of P-224, 20-40 corresponds to runs of B-233 etc. y axis shows the calculated Pearson coefficients.

The coefficient k_1 is represented by solid red dots, k_2 by blue crosses, k_3 by blue circle and k_4 by blue triangles. Fig. 4-(a) shows the analysis results for the not optimized multiplier without any reductions. Fig. 4-(b) shows the analysis results for the optimized multiplier with the reductions that was synthesized for 10 MHz clock frequency. Fig. 4-(c) shows the analysis results for the optimized multiplier that was synthesized for 71 MHz clock frequency.

Our field multiplier was realized without any countermeasures against SCA, i.e. we did not use any kind of randomization of the multiplication sequence and multiplications are not masked. Coefficients represented by red points are indistinguishable from other coefficients, i.e. our unified multiplier is inherently resistant against the performed attacks. As consequence we propose to design a unified field multiplier, for ECs $P-224$, $P-256$, $B-233$ and $B-283$, based on the 4-segment Karatsuba MM. Compared to a multiplier using the classical MM (CMM), the proposed multiplier is faster i.e. it needs only 9 clock cycles compared to 16 clock cycles needed by CMM when the operands are segmented in exactly the same way. The benefit of our approach is that the resulting multiplier is inherently resistant against horizontal attacks.

IV. CONCLUSIONS

In this paper we presented a unified multiplier supporting 4 different elliptic curves. Please note that these curves belong to two different types of Galois fields namely $GF(p)$ and $GF(2^n)$. Which type of Galois field is currently supported is defined by a select signal. The calculations in both fields are different, especially the summation of partial products, as in $GF(p)$ carry-bits need to be obeyed. In order to ensure fast execution of the multiplication in both cases we integrated the carry bit separation technique to speed-up the addition in $GF(p)$.

In our design we use the 4 segment Karatsuba multiplication formula. which leads to the fact that the operands of the partial multiplications are quite long. In addition this multiplication method ensures that all partial products always have different operands. This is the reason why our unified multiplier is inherently resistant against HCCA Attacks. In order to proof this fact we run 20 experiments per elliptic curve calculated the Pearson coefficients for 4 different types of multiplications. In none of the 80 test runs the coefficients allowed to separate two groups of multiplications. So, it is infeasible to reveal the private by applying the Horizontal Collision Correlation Analysis Attack.

ACKNOWLEDGMENT

This research has been funded by the Federal Ministry of Education and Research of Germany under grant number 03ZZ0527A.

REFERENCES

[1] Federal Information Processing Standard (FIPS) 186-4, Digital Signature Standard; Request for Comments on the NIST-Recommended Elliptic Curves: 2015. <http://dx.doi.org/10.6028/NIST.FIPS.186-4>

[2] Z. Dyka, D. Kreiser, I. Kabin, and P. Langendoerfer, "Flexible FPGA ECDSA Design with a Field Multiplier Inherently Resistant against

HCCA," in 2018 International Conference on ReConfigurable Computing and FPGAs (ReConFig), 2018, pp. 1–6.

[3] I. Kabin, Z. Dyka, D. Kreiser, and P. Langendoerfer, "Unified field multiplier for ECC: Inherent resistance against horizontal SCA attacks," in 2018 13th International Conference on Design Technology of Integrated Systems In Nanoscale Era (DTIS), 2018, pp. 1–4.

[4] A. Bauer, E. Jaulmes, E. Prouff, and J. Wild, "Horizontal Collision Correlation Attack on Elliptic Curves", in *SAC 2013*, pp. 553–570.

[5] Z. Dyka: Analysis and prediction of area- and energy-consumption of optimized polynomial multipliers in hardware for arbitrary $GF(2^n)$ for elliptic curve cryptography. Dissertation, 2012.

[6] Z. Dyka, P. Langendoerfer: *Area efficient hardware implementation of elliptic curve cryptography by iteratively applying Karatsubas method*. Proc. of the Design, Automation and Test in Europe (DATE 2005), 2005, Vol.3, pp: 70-75.

[7] IHP - Innovations for High Performance Microelectronics, <https://www.ihp-microelectronics.com/en/start.html>

[8] J. F. Wakerly, "Digital Design Principles and Practices (5th Edition)", ISBN-10. 013446009X (ISBN-13 9780134460093)

[9] "Mathematical routines for the NIST prime elliptic curves." [Online]. Available: <https://apps.nsa.gov/iaarchive/library/ia-guidance/ia-solutions-for-classified/algorithm-guidance/mathematical-routines-for-the-nist-prime-elliptic-curves.cfm>. [Accessed: 06-Jun-2019].

[10] B. Chevallier-Mames, Mathieu Ciet, and Marc Joye: *Low-cost solutions for preventing simple side-channel analysis: Side-channel atomicity*, IEEE Transactions on Computers, VOL. 53, No. 6, June 2004, p. 760-768

[11] Synopsys, PrimeTime <http://www.synopsys.com/Tools/>

Construction of Length and Rate Adaptive MET QC-LDPC Codes by Cyclic Group Decomposition

Usatyuk Vasilii, Egorov Sergey
 South-West State University
 Department of Computer Science, Kursk, Russia
 Email: L@Lcrypto.com, sie58@mail.ru

German Svistunov
 Omsk State Technical University
 Department of Computer Engineering, Omsk, Russia
 Email: g.v.svistunov@gmail.com

Abstract—We introduce a quasi-cyclic construction method for improving performance of length and rate adaptive Multi-Edge Type LDPC (MET-LDPC) codes below Block Error Rate (BLER) 10^{-5} . Proposed method allows to construct nested code families with a code length variability based on 5G eMBB modular lifting of Base Graph 2. Constructed codes are optimized by code distance and graph properties. Simulation results under 50 iterations of sum-product decoding over an AWGN channel with QPSK modulation are provided for comparing to 5G eMBB codes. It shows near the same performance for information length $k < 600$, 0.1 – 0.3 dB coding gain on $k > 600$ under rate 1/3, and coding gain around 0.3 dB under rate 1/5.

Index Terms—Length adaptation; Code distance; eMBB; modular lifting; Simulated annealing lifting; Quasi-cyclic; Multi-Edge Type LDPC; MET-LDPC; Cyclic group decomposition

I. INTRODUCTION

Low-density parity-check (LDPC) codes are taking its origin from the work of Gallager, [1], and further developed in the works of Tanner [2] and MacKay [3]. Multi-edge Type (MET) approach for LDPC codes is based on the idea of code-on-the-graph puncturing according to the special erasure recoverability distribution [4]. Code based on MET-approach requires more iterations for decoder convergence but it provides better iterative decoding threshold. In the LTE standard the Turbo code MET-approach was used for the improving of error-correcting properties by the cost of 6% of variable nodes punctured in circular buffers [5].

MET quasi-cyclic (QC) LDPC codes are currently applied at many modern standards: 5G eMBB [6], TV physical layer standard ATSC 3.0 [7], Deep Space Communication [8], fiber optic communication standard IEEE P802.3ca 50G-EPON [9]. MET QC-LDPC codes in 5G eMBB were constructed to provide the best performance at BLER 10^{-2} . That is why at BLER 10^{-5} it suffers from error-floor. In wireless communication systems a granularity of information length (transport block size) and compact representation of code specification are required. QC-LDPC codes allow to vary information length (length adaptation) by using different sizes of circulant, puncturing and shortening.

The key contribution of this work is a method for constructing the length and rate adaptive MET QC-LDPC codes under 5G by cyclic group decomposition modular lifting. Proposed

method is applying simulated annealing approach with the maximization of girth, EMD constrains and code distance optimization. A probabilistic Number Geometry method and Upper Bound of QC-LDPC code distance were used as well to improve code distance properties of the shortest codes.

This paper is organized as follows. In Section II we introduce MET QC-LDPC code definitions and notations. In Section III we describe a method for construction of the length adaptive MET QC-LDPC codes. In Section IV we describe a method for Simulated annealing based lifting with cycle (graph) properties from 5G eMBB protograph (BG2). A code distance improving for the short MET QC-LDPC codes was considered in section V. The performance of MET QC-LDPC codes based on proposed approach was investigated by simulations in Section VI.

II. MULTI-EDGE TYPE QC-LDPC CODES

A QC-LDPC code parity-check matrix is formed by circulant permutation sub-matrices (CPM) which could be either zero or nonzero matrices. Let $P = (P_{ij})$ be the $L \times L$ CPM defined as

$$P_{ij} = \begin{cases} 1, & \text{if } i + 1 \equiv j \pmod{L} \\ 0, & \text{otherwise.} \end{cases}$$

Thus CPM P^k which is obtained by shifting the identity matrix I to the right by k times for any k , $0 \leq k \leq L - 1$. In our notation we denote the zero matrix by P^∞ and the set $\{\infty, 0, 1, \dots, L - 1\}$ by A_L .

Let the $mL \times nL$ matrix H be defined as

$$H = \begin{bmatrix} P^{a_{11}} & P^{a_{12}} & \dots & P^{a_{1n}} \\ P^{a_{21}} & P^{a_{22}} & \dots & P^{a_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ P^{a_{m1}} & P^{a_{m2}} & \dots & P^{a_{mn}} \end{bmatrix},$$

where $a_{i,j} \in A_L$ and L is the circulant size of H .

Consider the *exponent matrix* of H denoted as $E(H) = (E_{ij}(H))$:

$$E(H) = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix},$$

i.e., the entry $E_{ij}(H) = a_{ij}$.

The *protograph mother matrix (base graph)* $M(H)$ is a $m \times n$ binary matrix obtained by replacing in H zero and nonzero CPMs by 0 and 1 respectively. For the reason of simplicity, we consider the case of edges of multiplicity one. However the proposed method and its implementation support multiple edges base graph [16].

Denote a cycle of length $2l$ in the Tanner graph of $M(H)$ as a *block-cycle* of length $2l$. Every block-cycle of length $2l$ corresponds both to the sequence of CPMs $\{P^{a_1}, P^{a_2}, \dots, P^{a_{2l}}\}$ in H and to the *exponent chain* $\{a_1, a_2, \dots, a_{2l}\}$ in $E(H)$.

The cycles in the parity-check matrix H Tanner graph could be found in the following manner [10]. An *exponent chain* forms a cycle in the Tanner graph of H if

$$\sum_{k=1}^{2l} (-1)^k a_k \equiv 0 \pmod{L}. \quad (1)$$

In equation (1) each coefficient $a_{i,j}$ for CPM $P^{a_{i,j}}$ is added with a plus for every even step and with a minus for every odd step. If the number of even steps for CPM $P^{a_{i,j}}$ is equal to the number of odd steps, then $a_{i,j}$ is eliminating from the equation.

Let's consider a cycle in the Tanner graph and a set VN_{cycles} of all the variable nodes involved in that cycle. The number of check nodes singly connected to the variable nodes from VN_{cycles} is a metric of the cycle in the Tanner graph named Extrinsic Message Degree (EMD) [13]. The EMD is a metric which shows the connection measure of the subgraph (cycle) with the rest of the Tanner graph.

III. LENGTH ADAPTATION OF MET QC-LDPC CODES

Consider a QC-LDPC code with $m \times n$ exponent matrix $E(\mathbf{H}_0) = (E_{ij}(\mathbf{H}_0))$ and a base graph $M(\mathbf{H}_0)$. A parity-check matrix H_0 of this code will be of size $mL_0 \times nL_0$ where L_0 is the circulant size. *Lifting* is a method of constructing QC-LDPC codes with $mL_k \times nL_k$ parity-check matrices \mathbf{H}_k , same base graph $M(\mathbf{H}_k) = M(\mathbf{H}_0)$ and entries of exponent matrices $E(\mathbf{H}_k)$ satisfy $E_{ij}(\mathbf{H}_k) \in \mathcal{A}_{L_k}$ from \mathbf{H}_0 for a given set of circulant sizes $\{L_k\}$, $L_k < L_0$. So it is possible to specify a formula for recalculating the values of $E(\mathbf{H}_k)$ from $E(\mathbf{H}_0)$. In [14] modular and floor-modular lifting approaches are given. Length adaptation methods theoretical analysis and its generalization were proposed at paper [15].

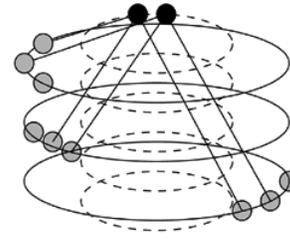
Modular lifting is determined by the following equation, $E(H_k) = f_{\text{modular}}(E(\mathbf{H})_i, k)$ [14]:

$$E_{ij}(\mathbf{H}_k) = \begin{cases} E_{ij}(\mathbf{H}_0) \pmod{L_k}, & \text{if } E_{ij}(\mathbf{H}_0) \neq -1, \\ -1, & \text{otherwise.} \end{cases} \quad (2)$$

To improve the properties of modular lifting in 5G standard 8 cyclic groups (different families) are used, Fig. 1:

$$L_k = a \cdot 2^i, \quad (3)$$

where $i = \{0, 1, 2, 3, 4, 5, 6, 7\}$ — set of index, $a = \{2, 3, 5, 7, 9, 11, 13, 15\}$ — code's family lifting values, L_k — set of lifting sizes, (see "Table 5.3.2-1" at [6]).



For m to n protograph

$$L_{256} = L_2 \times L_2 \times \dots \times L_2$$

$$L_{384} = L_3 \times L_2 \times \dots \times L_2$$

$$L_{320} = L_5 \times L_2 \times \dots \times L_2$$

$$L_{224} = L_7 \times L_2 \times \dots \times L_2$$

...

$$L_{240} = L_{15} \times L_2 \times \dots \times L_2$$

● code bit nodes (n rings of L_k each)

○ constraint nodes (m rings of L_k each)

Fig. 1: Modular lifting of QC-LDPC codes as cyclic group decomposition

IV. METHOD FOR CONSTRUCTION LENGTH AND RATE ADAPTIVE MET QC-LDPC CODES

Consider a family of MET QC-LDPC codes with $mL_k \times nL_k$ parity-check matrix \mathbf{H} where L_k are the circulant sizes, $m \times n$ exponent matrix $E(\mathbf{H})$ and base graph $M(\mathbf{H})$. $N = L_k \times n$ is a code length, $R = \frac{n-m-s}{n-p-s}$ is a code rate. Here n define the number of variable nodes in a base graph, m — number of check nodes in a base graph, p — number of punctured CPMs, s — number of shortened CPMs. Let we have Base graph 2 (BG2) base matrix (protograph) with size $m = 52, n = 42, p = 2$, code rate $1/5$ and nested submatrix with size $m = 32, n = 22, p = 2$ with code rate $1/3$. For the first step of our construction algorithm we generate 8 different exponent matrix for each family of cyclic group $E(\mathbf{H}) = \max_{\text{girth, EMD}} M(\mathbf{H})$ using simulated annealing method with maximization of girth and EMD for every circulant size $E(\mathbf{H}) = \text{SALift}(M(\mathbf{H}), L_k)$ [16]. The graph properties optimization allows us to construct codes achieving up to 0.3 dB gain under $FER = 10^{-5}$ compared with eMBB 5G code for information lengths from 600 to 1600, Fig. 2. But those codes still have 0.6 dB gap for information length below 232 in case of code rate $1/3$ (Fig. 3) and a gap below information length 380 in case of code rate $1/5$ (Fig. 4). To solve the short codes performance problem we shall search for short lengths QC extension of BG2 with better cycles properties (*girth, EMD*) and Hamming distance. Maximal performance gap for constructed codes in 5G standard cases observed for lengths which got decreased number of information symbols in mother matrix(protograph): from 10 to 6 variable nodes for transport block size below 192 and 8 for [192, 560], p. 10 [6]. Proposed method allow to improve code distance, e.g. for family $a = 2$, for CPM size 8 from 20 (5G) to 23, for CPM size 32 from 31 (5G) to 44. The sieving method for MET QC-LDPC codes by Hamming distance and (*girth, EMD*) optimization is described by the pseudo code, Alg.1.

V. NUMBER GEOMETRY BASED PROBABILISTIC CODE DISTANCE ESTIMATION METHOD

Code distance estimation could be performed by the Brower-Zimmerman method [17]. It was implemented in Magma algebraic system and GAP system for computational discrete

Algorithm 1 Codes Sieving Method For Construction of Length Adaptive MET QC-LDPC Codes

Require: $M(H)$ – mother matrix, L_0 -maximal lifting value, K –set of circulant sizes for code distance sieving, e.g. $\{4, 8, 16, 32\}$, $Card_c$ -number of lifted codes for sieving.

```

1:  $E_{sieve}(\mathbf{H}) = \emptyset$ 
2: for  $i = 0; i < Card_c; i = i + 1$  do
3:    $E(\mathbf{H})_i = SALift(M(\mathbf{H}), L_0)$ 
4:    $E(H_K) = f_{modular}(E(\mathbf{H})_i, K)$ 
5:    $dmin_{K,i} = NG(E(H_K))$ 
6: end for
   return  $E_{sieve}(\mathbf{H}) = \max_{dmin_{K,i}} E(H_K)$ 

```

where $E(\mathbf{H})_i$ - is the set of codes constructed by simulation annealing lifting $E(\mathbf{H}) = \max_{girth, EMD} M(\mathbf{H})$, $E(H_K)$ - is the length adapted codes obtained by using modular lifting for circulant size from set $\{K\}$, $dmin_{K,i}$ - is a code distance estimated by Alg. 2.

algebra. Code distance problem is a NP-hard and it is taking a lot of time even for structured QC-LDPC codes with relatively small circulant size $L_k = [16, 64]$. Fortunately for the code construction it is enough to use probabilistic algorithm and tight upper bound estimation which decrease the search time by several magnitudes comparing to the deterministic algorithm. Number Geometry probabilistic method is using Kannan's embedding technique and Lattice Construction A [18], [19]. Deterministic Number Geometry method for the code distance estimation was proposed at [20]. We generalize this approach to probabilistic one. Our generalization is based on the searching methods of the shortest vector in Ideal-Lattice applied for the Post-Quantum cryptography challenge [21], [22].

Geometry Lattice. Lattice – discrete Abelian subgroup defined under R^n . For the linear independent basis $B = \{b_1, \dots, b_n\}$ defined under R^m , lattice points are represented by linear integer combination:

$$L(b_1, \dots, b_n) = \left\{ \sum_{i=1}^n x_i b_i : (x_1, \dots, x_n) \in Z^n \right\},$$

where m and n are dimension and rank of lattice respectively, $m \geq n$.

Shortest Vector Problem. (Δ -short vector problem, $SVP_\Delta(m)$): For the defined m -dimension lattice $L(B)$ with rank n and real value $\Delta > 1$ we need to find a non-trivial vector, Δ -large shortest vector in the lattice $\bar{b} \in L : \|\bar{b}\| \leq \Delta \cdot \lambda_1(L)$.

For the case when $\Delta = 1$, we solve shortest vector problem. For the case when $\Delta > 1$, we solve short vector problem (approximation of SVP). At papers [23], [24] a deterministic method for this classical Number Geometry problem solution is proposed.

Shortest Basis Problem. (Δ -short basis problem, $SBP_\Delta(m)$): For defined full rank lattice basis B and real value $\Delta > 1$ we need to find a lattice basis $B' =$

$$\{b'_1, b'_2, \dots, b'_m\} : L(B) = L(B'), \prod_{i=1}^m \|b'_i\| \leq \Delta \cdot \prod_{i=1}^m \|b_i\|.$$

Length reduced basis. Consider lattice $L \subset R^m$. We denote a basis $B = \{b_1, b_2, \dots, b_m\}$ as length reduced, if for QR-decompose orthogonal basis following inequality holds:

$$|\mu_{i,j}| \leq \frac{1}{2}, 1 \leq j < i \leq m,$$

where $\mu_{i,j}$ is a Gramm-Shmidt coefficient.

Block Korkin-Zolotarev reduced basis [25]. Consider a basis $B = \{b_1, b_2, \dots, b_m\}$ of lattice $L \subset R^m$ ordered by length. We will denote it as reduced by Block Korkin-Zolotarev with block size $\beta \in [2, m]$ and precious $\delta \in (\frac{1}{2}, 1]$ if basis B is length reduced and: $\delta^2 \cdot \|b_i^\perp\|^2 \leq \lambda_1^2(L_i), i = 1, \dots, m$. Here $\lambda_1(L_i)$ is the length of shortest vector in L_i , defined by orthogonal complement of a vector space with a basis $b_i, \dots, b_{\min(i+\beta-1, m)}$.

Lenstra-Lenstra-Lovasz (LLL) reduced basis, [26]. Lattice basis called LLL-reduced if it's a Block Korkin-Zolotarev (BKZ-reduction) reduced basis with $\beta = 2$ precious $\delta \in (\frac{1}{2}, 1)$.

Deterministic Code distance estimation methods are based on Number Geometry [20]:

I. Consider a generator matrix of binary, ternary linear code G embedded into the geometrical lattice B_c . For this purpose we will scale $n - k$ subspace of basis B_c by a constant scale value N in a such way to map Hamming distance between codewords to Euclidean distance between lattice points. Basis of lattice is given as

$$B_c^T = \begin{pmatrix} N \cdot G & I_k \\ N \cdot q \cdot I_n & 0^{n \times k} \end{pmatrix},$$

where $G \in F_q^{k \times n}$ is a code's generator matrix on the alphabet of size q , I_k is an identity matrix, B_c^T - transposed basis of the lattice.

For the systematical codes $G = (I_k|P)$ and $P \in F_q^{k \times n-k}$ a scale constant N is equal to 1. The resulting lattice is not the full rank lattice but $rank(B_c) = k$.

II. Reduce the basis of this lattice by solving shortest basis problem.

III. Solve shortest vector problem in the lattice using Kannan-Finke-Post method [23], [24]. Number of non-zero positions in the shortest vector defines Hamming weight.

To realize a probabilistic algorithm Alg. 2 we will enumerate all integer x in $L: x \in \{L \cap S \cap P\}$, where S - sphere of radius defined by code distance upper bound A , P -subset of Lattice points which most probably contains a codeword of minimal weight. This task can be solved by probabilistic sampling as it shown in paper [27] with an assumption about QR-decomposed orthogonal basis length and some properties of the search area. More practical assumption was proposed at paper [28]. It was also generalized for any type of lattice subset at papers [29], [30].

To optimize the search area we propose to use random permutations of lattice basis and defining the search area P in such way to minimize the expectation Exp and dispersion Var of the search area points set cardinality. It allows to get shortest vector with defined probability. For example, for area P defined by inequality

$$P = \left\{ \sum_{i=1}^n x_i b_i^\perp : \frac{t_i}{2} < x_i \leq \frac{t_{i+1}}{2} \right\}, t \in \mathbb{N}^n.$$

Expectation and variance of the set are calculated as follows:

$$Exp = \sum_{i=1}^n \left(\frac{t_i^2}{4} + \frac{t_i}{4} + \frac{1}{12} \right) \|b_i^\perp\|^2,$$

$$Var = \sum_{i=1}^n \left(\frac{t_i^2}{48} + \frac{t_i}{48} + \frac{1}{180} \right) \|b_i^\perp\|^4.$$

Exp and Var could be calculated efficiently by QR-decomposition with using of Givens rotation on GPU or Householder reflection on multicore CPU. Generalization of Alg. 2 allows to solve binary and ternary Learning with Error Problem (LWE) which is equivalent to the Bounded Distance Decoding of code embedded into lattice [31], [32].

A generalization of Mackay-Davey Theorem 2 [33] could be used as an Upper Bound of code distance d_{min}^{upper} . It is known also as Smarandache-Vontobel upper bound of QC-LDPC code distance [34] improved by Butler-Siegel [35].

Theorem 12 [35]. Let C' be a QC code constricted by optionally puncturing subblocks of the QC code C , defined by the polynomial parity-check matrix $H(x) \in ((F_2[x]/\langle x^N - 1 \rangle)^{J \times L})$ and let $\triangleq wt(H(x))$. Let the subblock of C indexed by the set P , $P \subset [L]$, be punctured, while maintaining the dimensionality of the code. Let A' be a submatrix of A with rows a_t , $t \in \tau \subset [J]$, removed. Let S be a subset of $[L]$ of size $J + 1 - |\tau|$, such that the subrows $a_{t,S} = 0 \forall t \in \tau$. Then

$$d_{min}(C') \leq \min_{S,\tau} \sum_{i \in S \setminus P} perm(A'_{S \setminus i}).$$

The proposed method source code and constructed MET QC-LDPC codes available at [37].

Algorithm 2 Number Geometry based probabilistic code distance estimation method

Require: G —Code generator matrix, $Type$ —type of searching area, d_{min}^{upper} —upper bound on code distance, Num —number of random permutation of lattice basis, q —code alphabet $q \in 2, 3$, β —block size in BKZ-basis reduction method, δ —precious of length reduction.

1: Embedded code to lattice

$$B_c^T = \begin{pmatrix} N \cdot G & I_k \\ N \cdot q \cdot I_n & 0^{n \times k} \end{pmatrix}$$

2: $B' = SBP_\Delta(m)$ using BKZ(β, δ)

3: Generate Num permutation of basis B'

4: **for** $basis = 0; basis \leq Num; Num = Num + 1$ **do**

5: $QR(perm_{basis}(B))$

6: Exp_{basis}^{Type} and Var_{basis}^{Type}

7: **end for**

8: $B^* = \min(Exp_{basis}, Var_{basis})$

9: $c = SVP_\Delta L(B^*)$ with radius $R \leq d_{min}^{upper}$, area $Type$

return d_{min} number of non-zero position in c

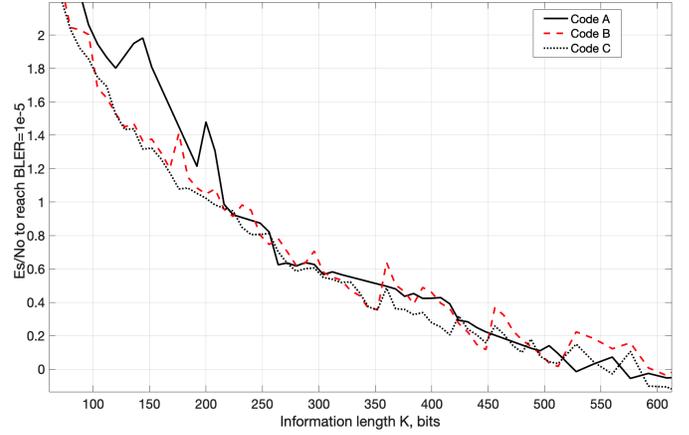


Fig. 2: SNR required for information lengths k to achieve FER level 10^{-5} for MET QC-LDPC codes with rate $1/3$.

VI. SIMULATION RESULTS

Performance of length adaptive MET QC-LDPC codes over an AWGN channel with QPSK modulation under the 50 iterations of sum-product decoding was analyzed by computer simulations. Figure 2, 3 shows the SNR value required to achieve 10^{-5} FER performance for information lengths $k = 80$ to 1600 for QC-LDPC codes with rate $1/3$, $1/5$. We considered 105 different lengths (from 80 to 512 with granularity 8, from 512 to 1024 with granularity 16 and from 1024 to 1600 with granularity 32) for every rate. Code A denotes the simulation results of MET QC-LDPC codes with 5G eMBB standard base graph 2 (BG2), [6]. Code B denotes the simulation results of 8 cyclic groups codes families constructed using Simulation Annealing lifting with maximization of girth and EMD. Code C denotes the performance curve of MET QC-LDPC codes sieved by the code distance.

VII. CONCLUSION

The proposed methods for construction of length adaptive MET QC-LDPC codes allows to improve cycle properties of BG2 codes and show near same performance for information length below 600 and from 0.1 to 0.3 dB gain on length greater 600 and rate $1/5$ compared to MET QC-LDPC codes from 5G.

REFERENCES

- [1] R. G. Gallager, "Low-Density Parity-Check Codes", Cambridge, 1963.
- [2] R. M. Tanner, "A recursive approach to low complexity codes," IEEE Trans. Inform. Theory, vol. 27, pp. 533-547, Sept. 1981.
- [3] D. MacKay, R. Neal, "Good codes based on very sparse matrices," Cryptography and Coding, Lect. Notes in Comp. Science, Oct. 1995.
- [4] T. J. Richardson and R. L. Urbanke, "Multi-edge type LDPC codes," in Workshop honoring Prof. Bob McEliece on his 60th birthday, California Institute of Technology, Pasadena, California, 2002.
- [5] J. Chenget et al., "Analysis of Circular Buffer Rate Matching for LTE Turbo Code," 68th Vehic. Techn. Confer., Calgary, BC, 2008, pp. 1-5.
- [6] 3GPP TS38.212V15.4.0:NR:Multiplexing and channel coding(Rel. 15)
- [7] K.-J. Kim et al., "Low-Density Parity-Check Codes for ATSC 3.0," IEEE Transactions on Broadcasting, 62(1), 189-196.
- [8] CCSDS 131.0-B-3 Standard, Issue 3 September 2017 Washington, DC, USA, <https://public.ccsds.org/Pubs/131x0b3e1.pdf>

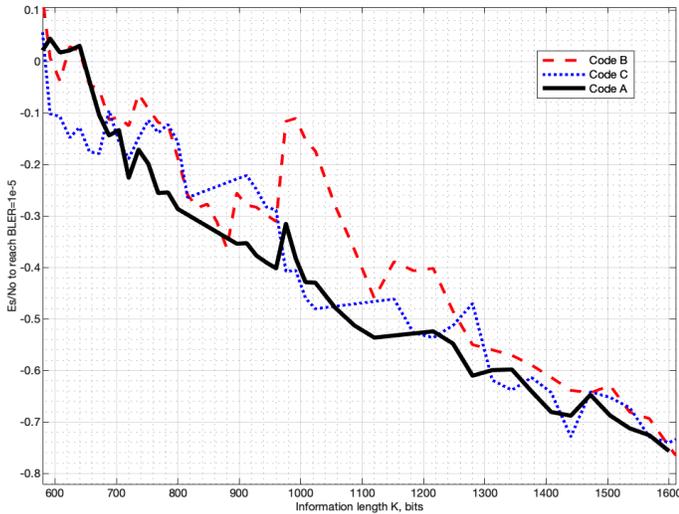


Fig. 3: SNR required for information lengths k to achieve FER level 10^{-5} for MET QC-LDPC codes with rate $1/3$.

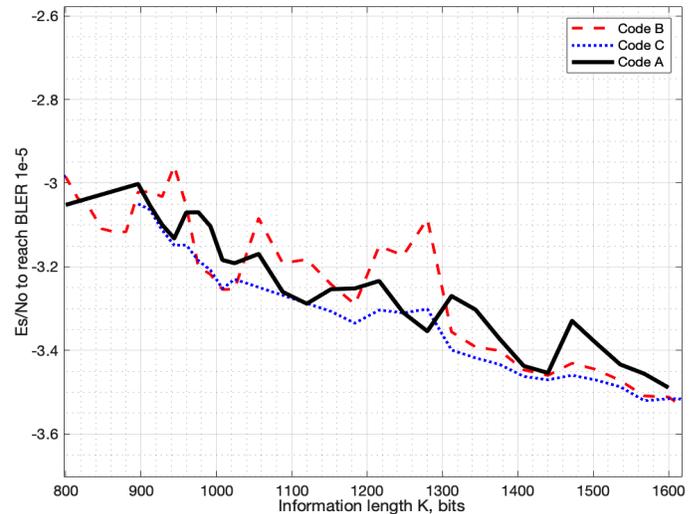


Fig. 5: SNR required for information lengths k to achieve FER level 10^{-5} for MET QC-LDPC codes with rate $1/5$.

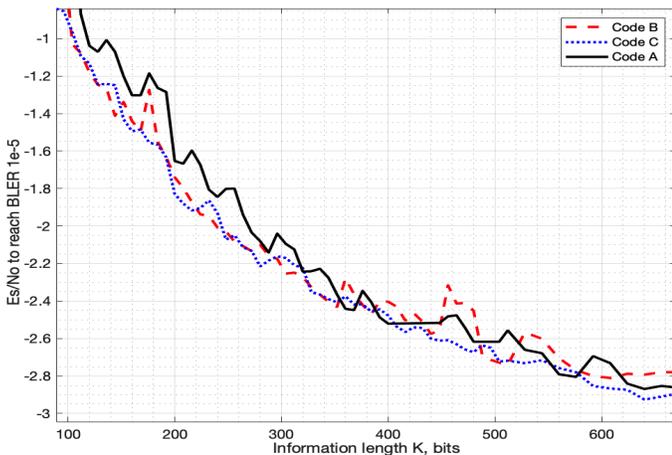


Fig. 4: SNR required for information lengths k to achieve FER level 10^{-5} for MET QC-LDPC codes with rate $1/5$.

[9] IEEE P802.3ca 50G-EPON, May 23-26, 2017, New Orleans, LA, USA
 [10] M.P.C. Fossorier, "Quasi-cyclic low-density parity-check codes from circulant permutation matrices," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1788–1793, 2004.
 [11] B. Vasic et al., "Trapping set ontology," 2009 47th Annual Allerton Confer. on Commun., Control, and Comput., 2009, pp. 1-7.
 [12] J. Chen, et al, "Improved min-sum decoding algorithms for irregular LDPC codes," *IEEE Inter. Symp. Inform. Theory*, 2005, pp. 449-453.
 [13] S. John, H. M. Kwon, "Approximate cycle extrinsic message degree regular quasi circulant LDPC codes," *IEEE Military Comm. Confer., Atlantic City, NJ, Vol. 5*, pp. 2877-2881, 2005.
 [14] S. Myung, K. Yang, "Extension of quasi-cyclic LDPC codes by lifting," *Proc. Intern. Symp. on Inform. Theory*, 2005., pp. 2305-2309.
 [15] I. Vorobyev, N. Polyanskiy et al., "Generalization of Floor Lifting for QC-LDPC Codes: Theoretical Properties and Applications," 2018 IEEE East-West Design & Test Symposium (EWDTS), Kazan, 2018, pp. 1-6.
 [16] V. Usatyuk, I. Vorobyev, "Simulated Annealing Method for Construction of High-Girth QC-LDPC Codes," 2018 41st International Confer. on Telecom. and Signal Processing (TSP), Athens, 2018, pp. 1-5.
 [17] M. Grassl, "Searching for linear codes with large minimum distance," In: W. Bosma, J. Cannon (eds) "Discovering Mathematics with Magma. Algorithms and Computation in Mathematics," v. 19., 2006, pp 287-313

[18] R. Kannan "Minkowski convex body theorem and integer programming," *Mathematics of operations research*, 12(3):415-440, 1987.
 [19] J.H. Conway, N.J.A. Sloane, "Sphere Packings Lattices and Groups," in New York, N.Y.:Springer-Verlag, 1988.
 [20] A. Betten et al "Error-Correcting Linear Codes Classification by Isometry and Applications," Berlin: Springer-Verlag, 2006. 818 p., pp. 594-596
 [21] T. Plantard, M. Schneider, "Creating a Challenge for Ideal Lattices," *IACR Cryptology Rep.*2013/039
 [22] V.S. Usatyuk, O.V. Kuzmin, "Parallel algorithms for integer lattices basis reduction," *Software&Systems.*, vol. 1(109), 2015, p. 55-62
 [23] R. Kannan, "Improved algorithms for integer programming and related lattice problems," In *Proc. 15th ACM STOC*, pages 193–206, 1983.
 [24] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Mathematics of Computation*, 44(170):463–471, 1985.
 [25] C. P. Schnorr, "Block Reduced Lattice Bases and Successive Minima," *Combinatorics, Probability and Computing.*, 1994.,v. 3, p. 507-522.
 [26] A. K. Lenstra, H. W. Lenstra Jr., L. Lovász, "Factoring polynomials with rational coefficients," *Mathem. Annalen*, 1982, 261 (4): 515–534.
 [27] C. Schnorr, "Lattice Reduction by Random Sampling and Birthday Methods," *STACS of Lect. Notes in Comp. Science*, Springer, 2003, pp. 145-156.
 [28] J. Buchmann, C. Ludwig, "Practical lattice basis sampling reduction," In *ANTS*, v. 4076 of LNCS, pp. 222-237. Springer, 2006.
 [29] A. Yoshinori, Phong Q. Nguyen, "Random Sampling Revisited: Lattice Enumeration with Discrete Pruning," *Cryptology Rep.* 2017/155
 [30] M. Yoshitatsu, T. Tadanori, K. Kenji, "Estimation of the Success Probability of Random Sampling by the Gram-Charlier Approximation," *Cryptology ePrint Archive*, Report 2018/815,2018, <https://eprint.iacr.org/2018/815>
 [31] E. Kirshanova et al. "Parallel Implementation of BDD Enumeration for LWE," *ACNS 2016. Lecture Notes in Computer Science*, vol 9696.
 [32] M. R. Albrecht et al, "The General Sieve Kernel and New Records in Lattice Reduction," *Cryptology ePrint Archive*, Report 2019/089
 [33] J. MacKay, M. C. Davey, "Evaluation of Gallager codes for short block length and high rate applications," *roc. of the IMA Workshop on Codes, System and Graphical Models*, 1999. Springer-Verlag 2001, pp. 113-130
 [34] R. Smarandache, P.O. Vontobel, "Quasi-cyclic LDPC codes: influence of proto- and Tanner-graph structure on minimum hamming distance upper bounds," in *IEEE Trans. Inf. Theory*, vol. 58(2), pp. 585-607, Feb. 2012.
 [35] B.K. Butler, P.H. Siegel, "Bounds on the minimum distance of punctured quasi-cyclic LDPC codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4584-4597, July 2013.
 [36] V.S. Usatyuk, Constructed codes, (girth,EMD) and code distance optimization <https://github.com/Lcrypto/Length-und-Rate-adaptive-code>
 [37] R. Andrew, N. Dingle, "Implementing QR Factorization Updating Algorithms on GPUs," *Parallel Comput.*, 2014, V. 40(7) pp. 161–172

A Technique for the Accounting of Surrounding Circuitry During Generation of the Simplified Models

Mark M. Gourary
CAD department
IPPM RAS
Moscow, Russian
Federation
gourary@ippm.ru

Sergey G. Rusakov
CAD department
IPPM RAS
Moscow, Russian
Federation
rusakov@ippm.ru

Sergey L. Ulyanov
CAD department
IPPM RAS
Moscow, Russian
Federation
ulyas@ippm.ru

Michael M. Zharov
CAD department
IPPM RAS
Moscow, Russian
Federation
zarov@ippm.ru

Abstract—The problem of decreasing the redundancy of model order reduction (MOR) techniques in circuit simulation is discussed. It is shown that further progress in decreasing the redundancy of the reduced models can be achieved by taking into account the characteristics of the circuit environment of the original models. Some experimental results are presented.

Keywords—analog integrated circuits, circuit simulation, model order reduction

I. INTRODUCTION

Model order reduction techniques (MOR) are widely used in practical design of integrated circuits [1-5]. The projection methods of MOR are based as a rule on Krylov subspace methods (see for instance [1-9]). However the redundancy is peculiar to the reduced models obtained by this way [9]. The redundancy means in this case that the order of the reduced model is often much larger than that required to provide the desired accuracy.

The redundancy problem becomes stronger for multipoint applications [8-10]. The existing projection based methods lead to a larger reduced model than it is necessary when the number of ports is large [5]. As mentioned in [5] “Krylov subspace projection methods will be extremely inefficient” for reducing the models with many ports. The extension of Krylov subspace methods to the multipoint case by applying block-Krylov techniques (see for instance [9, 12]) requires the deflation of redundant columns.

Multipoint projection methods [10, 11] produce more compact reduced models [14, 15]. The suggested method PMTBR allowed to connect multipoint techniques with truncated balance (TBR) algorithms [4, 13] and provided the theoretical basis for further development of multipoint direction in MOR techniques. New results in point selection for multipoint MOR were obtained in [16, 17]. These methods have some practical limitations, in particular the problem of error control in reducing process is open [10, 14]. The approach to generate reduced basis in multipoint projection under error control was developed in [18].

In this paper a new approach to decrease the redundancy of the reduced models is developed that is directed to take into account the characteristics of the circuit environment of the original models during model order reduction process.

II. PROJECTIVE TECHNIQUES IN MODEL ORDER REDUCTION

A. Description of Linear System

Let the model of a linear system with N_{inp} inputs, N_{out} outputs, N internal states be described in the following matrix form:

$$(G + sC)X = B; Z = D^T X, \quad (1)$$

where s is the Laplace variable. Matrices in the system (1) have the following orders: $G, C (N \times N), B (N \times N_{inp}), D (N \times N_{out}), Z (N_{out} \times N), X (N \times N_{inp})$.

The system (1) can be written in the form

$$X(s) = Y(s)^{-1} B, Z(s) = D^T Y(s)^{-1} B, \quad (2)$$

where $Y(s) = G + sC$

The moments of the system (1) solution $X^{[k]}(s)$ at frequency point s are obtained by successive differentiation (2) and have the view:

$$X^{[0]}(s) = X(s) = Y(s)^{-1} B, \\ X^{[k]}(s) = \frac{1}{k!} \frac{d^k}{ds^k} X(s) = -Y(s)^{-1} C X^{[k-1]}(s). \quad (3)$$

In the case of single input ($N_{inp} = 1$, B and $X^{[k]}(s)$ are vector-columns) the moments define Krylov subspace with matrix $-Y(s)^{-1} C$. For $N_{inp} > 1$ expressions (3) specify block Krylov subspace.

Projection methods of MOR provide transformation of the system (1) into a system of order N_{red} ($N_{red} < N$) using the multiplication of (1) by the right matrix $V (N \times N_{red})$ and left matrix $W (N \times N_{red})$:

$$(\hat{G} + s\hat{C})\hat{X} = \hat{B}, \hat{Z} = \hat{D}^T \hat{X}, \quad (4)$$

where

$$\hat{G} = W^T G V, \hat{C} = W^T C V, \hat{B} = W^T B, \hat{D} = V^T D. \quad (5)$$

Let i -th columns of matrices B, D, X, Z, V be denoted as $b^{(i)}, d^{(i)}, x^{(i)}, z^{(i)}, v^{(i)}$.

The solution of the system (1) at point s for the unit i -th input can be approximated by a linear combination of columns of the matrix V ($v^{(k)}$) with coefficients that are variables of the reduced system (4):

$$x^{(i)}(s) \approx \tilde{x}^{(i)}(s) = V \cdot \hat{x}^{(i)}(s) = \sum_{k=1}^{N_{red}} \hat{x}_k^{(i)}(s) v^{(k)}, \quad i=1,2,\dots,N_{inp}. \quad (6)$$

Here $\tilde{x}^{(i)}(s)$ is the approximation of $x^{(i)}(s)$, $\hat{x}_k^{(i)}(s)$ is k -th element of the vector.

The residual vector of the system (1) after substitution (6) must be orthogonal to all columns of the matrix W . This gives the equations for columns (4), (5). The solution of the reduced system (4) at the point s is the coefficients of the linear form of the approximation of the solution (6).

The following statement can be easily indicated: the state vector of the initial system at some point s is a linear combination of columns of the matrix V if and only if this vector coincides with its approximation at this point.

This statement is true for higher order moments. The methods of moment matching are based on this statement [8]. In this case the projective matrix V is formed by constructing an orthonormal basis for a given set of moments.

If both projective matrices are equal ($W = V$) then the transformation (5) ensures the symmetry of matrices C, G and the passivity of the system [7].

One possible approach to decrease the redundancy of low-order models generated by Krylov subspace methods has been presented [18]. This State Vector Selection (SVS) approach provides the ability to control the error estimate. The proposed technique to take into account the circuit environment in the reducing process is based on this approach. Therefore its description is given below in brief form.

B. State Vector Selection Approach

The proposed algorithm with explicit computation of the projective matrix can be outlined as follows.

Preliminary step.

Solve the original system (1, 2) for the given set of Laplace points $s \in S = \{s_1, s_2, \dots, s_m, \dots\}$;

save the state vector $x(s_m)$ and the output vector $z(s_m)$ for each point s_m .

Numerical cycle.

Let $n-1$ basis vectors (columns of the current projection matrix $V^{(n-1)}$) be already obtained, the matrices of the current

reduced system of $(n-1)$ -th order have been evaluated by (5). Then the new basis vector is determined by the following steps.

1) Solve the current reduced system at all given Laplace points

$$\hat{x}^{(n-1)}(s_m) = \left(\hat{A}^{(n-1)}(s_m) \right)^{-1} \cdot b^{(n-1)}. \quad (7)$$

$$\hat{z}^{(n-1)}(s_m) = d^{(n-1)} \cdot \hat{x}^{(n-1)}(s_m). \quad (8)$$

2) Evaluate the maximal error at each sample point

$$e^{max}(s_m) = \max_{u \in U} \left\| \left(Z(s_m) - \hat{Z}^{(n)}(s_m) \right) \cdot u \right\|. \quad (9)$$

3) Select the worst-case sample point s_M with the maximal error norm:

$$M = \arg \max_m e^{max}(s_m). \quad (10)$$

4) If $e^{max} < E$ then the reducing cycle is finished. Here E is the specified upper bound of the error.

5) Determine the worst-case state vector $x^* = x(s_M)$.

6) Define a new basis vector by orthogonalization of state vector with respect to previous basis vectors

$$v^{(n)} = x^* \perp V^{(n-1)}, \quad V^{(n)} = [V^{(n-1)}, v^{(n)}]. \quad (11)$$

7) Compute matrices of the reduced system of the current order n by (7).

8) Go to step 1 with $n=n+1$.

At the first step no reduced system exists, so the reduced TF can be assumed to be zero $\hat{z}^{(0)}(s) \equiv 0$.

An important advantage of the SVS algorithm is that the error norm is not specified in advance in the form of any mathematical definition, but it can be flexibly changed depending on the user's requirements.

C. Models of Linear Electrical Networks for projective MOR methods

For the application of above mentioned methods to the reduction of linear electrical circuit models it is necessary to represent models equations in the form of (1).

If a circuit is described by the modified nodal analysis (MNA) [19], then the state variables x are the nodal voltages and inductance currents, and C, G are corresponding MNA matrices.

In this case, the full Laplace conductances of the matrix elements ($Y_{ij} = G_{ij} + s \cdot C_{ij}$) are related to the parameters of the linear components of the circuit ($y_{ij} = g_{ij} + s \cdot c_{ij}$) by the relations

$$Y_{ij} = -y_{ij} \quad (i \neq j), \quad Y_{ii} = \sum_{k=0}^N y_{ik}. \quad (12)$$

where y_{i0} is the conductance between the i -th node of the circuit and the “ground”.

To generate solutions $X(s_m)$, $Z(s_m)$ during the process of circuit models reducing unit current and/or voltage sources [7] are connected to external terminals of a circuit. If the i -th input signal is given by a current source connected to the k -th node of the circuit then element $B_{ki}=1$ is formed in the matrix B . If the circuit contains a voltage source the current of a voltage source becomes an additional (m -th) state variable, the matrix G is expanded by m -th column and row element B_{mi} is set up in the matrix B

$$\begin{aligned} G_{km} = G_{mk} = 1, \quad G_{jm} = G_{mj} = 0 (j \neq k) \\ B_{mi} = 1, \quad B_{mj} = 0 (j \neq i) \end{aligned} \quad (13)$$

Output variables are defined in a dual way as voltages of the current sources and currents of the voltage sources. As a result $D=B$.

III. MODEL ORDER REDUCTION OF LINEAR NETWORKS SUBCIRCUITS

The output signals of the reduced circuit are formed by the interaction of its components and components of circuit environment. To take into account such an interaction it is proposed to consider the task of reducing the circuit model within the framework of linear circuit environment approximately reflecting the operation conditions of this circuit.

The output resistances of signal sources and capacitive load are the examples of such a linear environment.

In further it is assumed that the interaction of the environment with the circuit is carried out only through the terminal nodes of the circuit which fully or partially coincide with the nodes of the circuit environment.

The problem of reducing the linear subcircuit model can be formulated by the following way. Let the given linear circuit be described by the state vector of dimension N . The part of this circuit is the reduced subcircuit with internal state vector x_I of dimension N_I . The circuit environment is described by the state vector x_S of dimension N_S . In this case the full state vector x consist of partial vectors: $x = [x_I \ x_S]^T$, $N = N_I + N_S$.

Indexes I and S denote the set of internal nodes and the set of environment nodes respectively.

The state of the circuit is determined from the system of Laplace equations

$$\left(\begin{bmatrix} G_{II} & G_{IS} \\ G_{SI} & G_{SS} \end{bmatrix} + s \begin{bmatrix} C_{II} & C_{IS} \\ C_{SI} & C_{SS} \end{bmatrix} \right) \cdot \begin{bmatrix} X_I \\ X_S \end{bmatrix} = \begin{bmatrix} 0 \\ B_S \end{bmatrix}. \quad (14)$$

Matrix indexes in (14) determine the set of nodes between which the corresponding elements of the conductance and capacitance matrix are given. The matrix B_S defines the signals amplitudes of N_{inp} voltage and current sources contained in the circuit environment.

As mentioned above it is assumed that the impact on the subcircuit is carried out only through the nodes of the circuit environment. The output characteristics are determined by the state vector of the circuit environment X_S which is the result of the solution (14).

The problem of reducing the subcircuit within a given circuit environment is defined here as a generation of matrices \hat{G}_{II} , \hat{C}_{II} , \hat{G}_{IS} , \hat{C}_{IS} , \hat{G}_{SI} , \hat{C}_{SI} of the reduced subcircuit. Herewith the proximity of the state of the complete initial circuit (14) to the state of the circuit with a reduced subcircuit model should be provided.

$$\left(\begin{bmatrix} \hat{G}_{II} & \hat{G}_{IS} \\ \hat{G}_{SI} & G_{SS} \end{bmatrix} + s \begin{bmatrix} \hat{C}_{II} & \hat{C}_{IS} \\ \hat{C}_{SI} & C_{SS} \end{bmatrix} \right) \cdot \begin{bmatrix} \hat{X}_I \\ \hat{X}_S \end{bmatrix} = \begin{bmatrix} 0 \\ B_S \end{bmatrix}. \quad (15)$$

In this case the matrices G_{SS} , C_{SS} of circuit environment are not reduced.

To solve this problem the Galerkin method is used, but in contrast to the reduction of the linear system (1) $N_S * N_I$ projective matrix is applied to approximate only the internal state of the subcircuit $x_I(s) \approx \tilde{x}_I(s) = V \cdot \hat{x}_I(s)$. The environment of the subcircuit must not be changed, i.e. $x_S(s) \approx \tilde{x}_S(s) = \hat{x}_S(s)$.

Then the approximation of the system (14) state based on the solution (15) is written as:

$$\begin{bmatrix} x_I(s) \\ x_S(s) \end{bmatrix} \approx \begin{bmatrix} \tilde{x}_I(s) \\ \tilde{x}_S(s) \end{bmatrix} = \begin{bmatrix} V \cdot \hat{x}_I(s) \\ \hat{x}_S(s) \end{bmatrix}. \quad (16)$$

Using (16) the residual of the system (14) is determined by the expression:

$$\begin{bmatrix} r_I \\ r_S \end{bmatrix} = \left(\begin{bmatrix} G_{II}V & G_{IS} \\ G_{SI}V & G_{SS} \end{bmatrix} + s \begin{bmatrix} C_{II}V & C_{IS} \\ C_{SI}V & C_{SS} \end{bmatrix} \right) \cdot \begin{bmatrix} \hat{x}_I \\ \hat{x}_S \end{bmatrix} - \begin{bmatrix} 0 \\ b_S \end{bmatrix}. \quad (17)$$

To uniquely identify a solution the orthogonality condition of the internal residual vector of the subcircuit to all columns of the projective matrix is added: $V^{-T} r_I = 0$. A zero residual value is also set up for all nodes in the environment: $r_S = 0$.

After performing these operations, one can obtain from (17) a linear system for \hat{x} :

$$\left(\begin{bmatrix} V^T G_{II}V & V^T G_{IS} \\ G_{SI}V & G_{SS} \end{bmatrix} + s \begin{bmatrix} V^T C_{II}V & V^T C_{IS} \\ C_{SI}V & C_{SS} \end{bmatrix} \right) \cdot \begin{bmatrix} \hat{x}_I \\ \hat{x}_S \end{bmatrix} = \begin{bmatrix} 0 \\ b_S \end{bmatrix}. \quad (18)$$

The obtained system (18) coincides with (15) at the following values of reduced matrices

$$\begin{aligned} \hat{G}_{II} = V^T G_{II}V, \quad \hat{G}_{SI} = G_{SI}V, \quad \hat{G}_{IS} = V^T G_{IS}, \quad \hat{G}_{SS} = G_{SS}, \\ C_{II} = V^T C_{II}V, \quad \hat{C}_{SI} = C_{SI}V, \quad \hat{C}_{IS} = V^T C_{IS}, \quad \hat{C}_{SS} = C_{SS}. \end{aligned} \quad (19)$$

Note that the elements of the matrices Y_{SI} , Y_{IS} can be nonzero only for N_B boundary (B) nodes of the subcircuit which are common to the subcircuit and its environment. Therefore, one can write

$$Y_{SI} = G_{SI} + sC_{SI} = \begin{bmatrix} Y_{BI} \\ 0 \end{bmatrix}, Y_{IS} = G_{IB} + sC_{IS} = \begin{bmatrix} Y_{IB} & 0 \end{bmatrix}. \quad (20)$$

Then the matrices of the subcircuit for all internal and boundary nodes can be reduced by the formulas obtained from (19, 20)

$$\hat{Y}_{II} = V^T Y_{II} V, \hat{Y}_{BI} = Y_{BI} V, \hat{Y}_{IB} = V^T Y_{IB}, \hat{Y}_{BB} = Y_{BB}. \quad (21)$$

The last equation in (21) shows that the matrices between the subcircuit terminals (BB , CB) are not changed during model reduction process.

Matrices of the reduced subcircuit model ($\hat{Y}^{sub} = \hat{G}^{sub} + s\hat{C}^{sub}$) combine components related to all nodes

$$\hat{G}^{sub} = \begin{bmatrix} \hat{G}_{II} & \hat{G}_{IB} \\ \hat{G}_{BI} & \hat{G}_{BB} \end{bmatrix}, \hat{C}^{sub} = \begin{bmatrix} \hat{C}_{II} & \hat{C}_{IB} \\ \hat{C}_{BI} & \hat{C}_{BB} \end{bmatrix}. \quad (22)$$

To simulate an electrical circuit with matrices (22), it is necessary to form conductances of the subcircuit components using (12, 22) and assign to the boundary nodes of the subcircuit the names of those nodes of the external circuit to which they are connected.

Note that the properties of subcircuit reduction (15, 19) are similar to the properties of reduction of standard form of linear models (4, 5) because expressions (21, 22) can be represented in an equivalent form:

$$\begin{aligned} \hat{G}^{sub} &= \bar{V}^T G^{sub} \bar{V} \\ \hat{C}^{sub} &= \bar{V}^T C^{sub} \bar{V} \end{aligned} \quad (23)$$

where $\bar{V} = \begin{bmatrix} V & 0 \\ 0 & E_B \end{bmatrix}$, E_B is a unit matrix of dimension

$N_B \times N_B$, transformation (21, 23) ensures the preservation of passivity.

In practice, the reduced circuit never works within fixed linear environment. The different approaches are possible to take into account an arbitrary circuit environment.

The external part of the circuit can be presented by several variants of the linear environment. In this case to take into account such a multivariance it is possible to exploit finding the maximum error not only for different value points of the input vector but also for the variants of the circuit environment.

It is also possible to specify the configuration of the external environment by parameters within the given limits. In this case, the evaluation of the sensitivity of the output characteristics to the environment parameters can reduce the

computational cost. Having obtained the sensitivity the maximum error can be estimated.

The solution of the problem for a given range of parameters of the external environment can serve as the basis for solving the problem of reducing the subcircuit with nonlinear environment. Nonlinear components can be represented, for example, by linear elements with values from the corresponding parameter range.

IV. NUMERICAL EXPERIMENTS

The multistage RC circuit (Fig.1) was selected to verify the workability of the proposed reducing algorithm with taking into account the characteristics of the circuit environment. The circuit contains 50 π type sections with $R=0.02$ Ohm, $C=0.05$ pF. The circuit environment includes the input unit voltage source ($V = 1V$) with internal resistance $R_S = 0.01$ Ohm and capacitive load, which varies about the average value of $C_L = 10pF$.

Two reduction techniques were compared in the numerical experiment: without taking into account the load ($C_L = 0$) and with the setting up the load value $C_L = 10pF$. The error tolerance $tol = 0.05$ V was specified. This tolerance means 5% deviation from the static transfer function.

As a result of applying MOR techniques the following conductance and capacitance matrices were obtained: matrices of order 4×4 for the case with taking into account the load and 6×6 for the case without taking into account the load. The increased order in the last case is explained by the expansion of the operating frequencies band of transfer function (TF). It can be seen from Fig. 2a.

Then the load $C_L = 10pF$ was connected to each model. The deviations of their TF from the TF of the full model with the same load were obtained. These deviations are shown in Fig. 2b. The computed plots show that taking into account the load during model order reduction process allows us to decrease resulting error approximately by two times under lower order of the generated models. More detailed comparison of reduced models depending on the specified error tolerance is presented in Table. 1.

The values of model orders and TF errors at a nominal load of 10pF are given in Table 1 for specified tolerance values. The column 4 shows the actual error achieved when reducing is performed without load account. The last column shows the ratio of the model errors for model reduction without taking into account the load and with taking into account the load.

It can be seen from this table that the gain from model generation with taking into account the real load is essentially growing with increasing the requirements for model accuracy.

Table 1 presents the results for the case of coincidence of the load value with the value used in reducing model order. In practice the average load can usually be known. Its value at any time depends on the nonlinearity of the circuit, mode changes, etc. So the numerical experiments were carried out to estimate the accuracy of the obtained models at a load deviation of $\pm 40\%$ from the nominal value. The results of these numerical experiments showing the dependence of the

ratio of the model errors on the load at different specified tolerances are given in Table 2. The significant efficiency growth of the suggested approach can be noticed with increasing the requirements to accuracy of the reduced model.

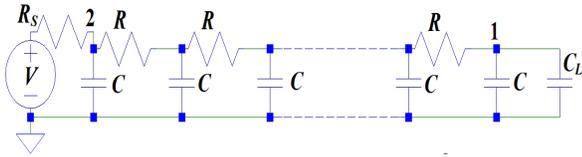


Fig. 1. RC multistage circuit with capacitive load.

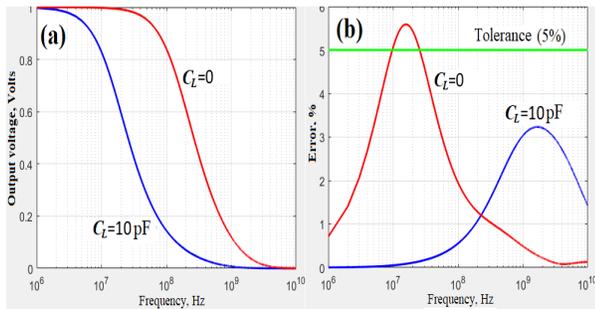


Fig. 2. Transfer functions of RC- circuit with different values of capacitive load (a), resulting errors versus frequency after reducing for different values of capacitive load (b).

V. CONCLUSION

The problem of reducing the order of the linear network subcircuit model taking into account the circuit environment is formulated. The basis for the system of Kirchhoff equations is formed in this case from the state variables vectors of the subcircuit obtained by solving the equations of the full circuit.

The model of the subcircuit is obtained in the form of matrices of capacitances and conductances and it is included into the circuit environment without the addition of auxiliary current or voltage sources.

Numerical experiments show a significant decreasing the error of model reduction by taking into account the average value of the load capacity.

REFERENCES

[1] W.H.A. Schilders, "The need for novel model order reduction techniques in the electronics industry," in *Model Reduction for Circuit Simulation*, Series: Lecture Notes in Electrical Engineering, vol. 74, Berlin: Springer-Verlag, 2011, pp. 3-23.

[2] W.H.A. Schilders, H.A. van der Vorst, J. Rommes, "Model order reduction: theory, research aspects and applications," in *Mathematics in Industry*, vol. 13, Berlin: Springer-Verlag, 2008.

[3] P. Benner, M. Hinze, E. J. W. ter Maten (Editors), *Model Reduction for Circuit Simulation*, Series: Lecture Notes in Electrical Engineering (LNEE). Springer, vol. 74, 2011.

[4] P. Benner, "Numerical linear algebra for model reduction in control and simulation," *GAMM Mitteilungen*, vol. 29, No. 2, pp. 275-296, 2006.

[5] S. X.-D. Tan, Z. Qi, and H. Li, *Advanced Model Order Reduction Techniques in VLSI Design*, Cambridge University Press, UK, 2007.

[6] P. Feldmann and R. W. Freund, "Efficient linear circuit analysis by Pade approximation via the Lanczos process," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 14, No. 5, pp. 639-649, 1995.

[7] A. Odabasioglu, M. Celik, L.T. Pileggi, "PRIMA: passive reduced-order interconnect macromodeling algorithm," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 17, No. 8, pp. 645-654, 1998.

[8] E. Grimme, "Krylov projection methods for model reduction," Ph.D thesis, Coordinated-Science Laboratory, Univ. of Illinois at Urbana-Champaign, Urbana-Champaign, IL, 1997.

[9] P.J. Heres, "Robust and efficient Krylov subspace methods for model order reduction," Ph.D thesis, Tech. Univ. Eindhoven, The Netherlands, 2005.

[10] I. M. Elfadel and D. L. Ling, "A block rational Arnoldi algorithm for multipoint passive model-order reduction of multiport RLC networks," in *Proc. Int. Conf. Comput.-Aided Des.*, San Jose, CA, Nov. 1997, pp. 66-71.

[11] N. Marques, M. Kamon, J. White, L. M. Silveira, "A mixed nodal-mesh formulation for efficient extraction and passive reduced-order modeling of 3D interconnects," in *Proc. 35th ACM/IEEE Des. Automation Conf.*, San Francisco, June 1998, pp. 297-302.

[12] P. Feldmann and R.W. Freund, "Reduced-order modeling of large linear subcircuits via a block Lanczos algorithm," in *Proc. 32nd ACM/IEEE Des. Automation Conf.*, San Francisco, June 1995, pp. 474-479.

[13] B.C. Moore, "Principal component analysis in linear systems: Controllability, observability and model reduction," *IEEE Trans. Autom. Control*, vol. AC-26, no. 1, pp. 17-32, 1981.

[14] J. R. Phillips and L. M. Silveira, "Poor man's TBR: A simple model reduction scheme," *IEEE Trans. Computer-Aided Design*, vol. 24, no. 1, pp. 43-55, Jan. 2005.

[15] J. R. Phillips, Z. Zhu and L. M. Silveira, "PMTBR: A family of approximate principal-components-like reduction algorithms," in *Model Reduction for Circuit Simulation*, Series: Lecture Notes in Electrical Engineering, vol. 74, Berlin: Springer-Verlag, 2011, pp. 111-132.

[16] L. M. Silveira and J. R. Phillips, "Resampling plans for sample point selection in multipoint model order reduction," *IEEE Trans. Comput.-Aided Des.*, vol. 25, no. 12, pp. 2775-2783, Dec. 2006.

[17] J. F. Villena and L. M. Silveira, "Multi-Dimensional Automatic Sampling Schemes for Multi-Point Modeling Methodologies," *IEEE Trans. Comput.-Aided Des.*, vol. 30, no. 8, pp. 1141-1151, Aug. 2011.

[18] M. M. Gourary, S. L. Ulyanov, M. M. Zharov "Model order reduction by state vector selection (SVS) approach," *European Conf. on Circuit Theory and Design (ECCTD)*, 2013, pp. 1-4.

[19] C. Ho, A. Ruehli, P. Brennan "The Modified Nodal Approach to Network Analysis," *IEEE Trans. on circuits and systems*, vol. 22, no. 6, 1975, pp. 504-509.

TABLE I. COMPARISON OF ERRORS FOR REDUCED MODELS AT THE NOMINAL CAPACITIVE LOAD (10 PF).

#	Tolerance	Reducing Conditions					Errors Ratio Err 0/Err C
		Without Load ($C_L=0$)			Load 10pF		
		Order	Deviation	Error	Order	Error	

1	2	3	4	5	6	7	8
1	10^{-1}	5	$7.7*10^{-2}$	$7.4*10^{-2}$	4	$3.2*10^{-2}$	2.30
2	10^{-2}	7	$9.8*10^{-3}$	$4.6*10^{-2}$	5	$5.9*10^{-3}$	7.68
3	10^{-3}	9	$2.1*10^{-4}$	$1.4*10^{-2}$	7	$4.8*10^{-4}$	30.1
4	10^{-4}	10	$1.9*10^{-5}$	$1.1*10^{-2}$	9	$1.3*10^{-5}$	867.5
5	$10^{-5}, 10^{-6}$	11	$7.0*10^{-7}$	$8.0*10^{-3}$	10	$6.8*10^{-7}$	11753
6	10^{-7}	12	$3.1*10^{-8}$	$6.2*10^{-3}$	11	$4.3*10^{-8}$	145110
7	10^{-8}	13	$2.3*10^{-9}$	$4.9*10^{-3}$	12	$2.2*10^{-9}$	2239200

TABLE II. COMPARISON OF ERRORS FOR REDUCED MODELS AT DEVIATION FROM THE NOMINAL CAPACITIVE LOAD.

#	C _L pF	Error Tolerance in Model Order Reduction						
		10^{-1}	10^{-2}	10^{-3}	10^{-4}	$10^{-5}, 10^{-6}$	10^{-7}	10^{-8}
1	14	3.2	10.9	42.7	1233	11309	17153	21650
2	12	2.8	9.27	36.4	1050	14141	25673	32399
3	10	2.3	7.68	30.1	867.5	11753	145110	2239200
4	8	1.8	6.09	23.8	685.1	7911	12025	15180
5	6	1.4	4.51	17.5	503.9	2338	3549	4480

A Signal Processing Approach for the Failure Analysis of Rolling-Element Bearing of Vehicle Brake Tester Used at a Vehicle Inspection Station

Selman Kulaç

*Department of Electrical-Electronics Engineering
Faculty of Engineering, Duzce University
Duzce, TURKEY
selmankulac@duzce.edu.tr*

Abstract—In this study, failure analysis of rolling-element bearing of vehicle brake tester used in a vehicle inspection station was considered according to the real time vibration signals. Based on vibration measurements under loads taken at intervals of about 15 days, these real time vibration signals were subjected to median filtering firstly. The total difference method proposed with this paper using autocorrelation functions' maximum and minimum values for each vibration signal was applied. Lastly, Goertzel algorithm has been used in order to obtain amplitude of specific defect frequency sample with a minimum calculation complexity for vibrational failure analysis of rolling-element bearing for the first time with this paper. It is verified with this paper that as the time passes, the values obtained by all methods above increase and excess the thresholds determined by this paper.

Index Terms—Failure analysis, rolling-element bearing, vibration measurement, vibration signal processing, vehicle roller brake tester, signal processing

I. INTRODUCTION

Vehicle braking systems are controlled periodically for their decisive importance in vehicle safety. The control of vehicle brakes are determined by the vehicle roller braking testers at vehicle inspection stations. Drum cylinders used in vehicle brake testers are mounted on rolling element bearings. Rolling element bearing fault is one of the most common causes of vehicle brake tester malfunctions. Therefore, failure analysis of rolling-element bearings used in a vehicle brake testers is necessary.

A considerable amount of literature has been published on signal processing based vibration monitoring and interpretations in order to determine bearing life and faults. Failure detection and diagnosis for a class of rolling-element bearings using short time FFT (Fast Fourier Transform)-based methods were investigated by Yang et al. [1]. Special auto-correlation analysis in frequency domain with an aim to distinguish the faults in the bearings was accomplished by Ming et al. [2]. The effect of sensor noise is noticeable in that study. In [3], vibration analysis in monitoring rolling-element bearings was

summarized and their capabilities, advantages and disadvantages were explained. In [4], the comparable efficiency of using frequency domain approach in HilbertHuang Transform for bearing fault diagnosis was handled. But empirical mode decomposition process existing in this study causes mode mixing problem. In [5], as the time passes, it is stated that the sum of the frequency amplitude values above the certain thresholds in frequency domain increases for the rolling-element bearing of a vehicle brake tester in loaded and unloaded cases. In [6], explicit finite element modelling of a rolling element bearing with a localised line spall using standard signal processing techniques etc. is obtained and analytical validation of the modelled results is presented. In [7], a novel modulation signal bispectrum (MSB) based robust detector for bearing fault detection which is more accurate and robust detection results than Kurtogram based approaches is proposed.

Although rolling element bearings have been widely used on brake testers at vehicle inspection stations, the bearing problems have not been considered before. In this study, a signal processing approach for the failure analysis of rolling-element bearings used at a vehicle inspection station was considered according to the real time vibration signals.

Novelties of this study are presented below.

In experimental :

- Taking the real time vibration signals from vehicle inspection station was performed.

In theory:

- The total difference method proposed with this paper for the first time using autocorrelation functions' maximum and minimum values for each vibration signal was applied for vibrational fault diagnosis of rolling-element bearing.
- Goertzel algorithm has been used in order to obtain amplitude of specific defect frequency sample with a minimum calculation complexity for vibrational fault diagnosis of rolling-element bearing for the first time with this paper.
- It is verified with this paper that as the time passes, the values obtained by all methods (total difference



Fig. 1. Pictures from experimental setup [5]



Fig. 2. Used bearing sample

based autocorrelation and Goertzel algorithm) increase and excess the thresholds determined by this paper.

II. REAL TIME VIBRATION SIGNALS OBTAINING AND ANALYSIS

The picture of the experimental setup is shown in Fig. 1 (a) and data acquisition (DAQ) device having accelerometer mounted on the bearing housing is shown in Fig. 1 (b). The bearing type is UCP205 and samples of healthy and broken down bearings are shown in Fig. 2. Two rolling element bearings are used for comparison. The first one was used to take measurement 1, 2 and 3 data and the other one was used to take measurement 4 and 5 data for comparison. Vibration measurements for bearings of vehicle roller brake testers were performed at a vehicle inspection station for intervals of about 15 days. Five measurements were carried out on experimental setup on different dates. The first bearing had been mounted on the brake test system at the vehicle inspection station before first measurement and was in standard active use. The first measurement was taken while active use was in progress. The measurements were then continued at regular intervals and the first bearing was broken about 35 days after the first measurement. It should be noted that the life of the first bearing is longer than one month.

Median filtering is a non-linear method for removing noise from images. It is widely used because it is very effective at removing noise and outliers while preserving edges in image processing. From Fig. 3 and Fig. 4, the expected peak values are observed approximately two in one second depending on the rotational speed of the bearing. The rotational speed of

the bearing in Hz is calculated by Eq. (1). The median filter was used instead of the mean filter to reduce noise without touching these peak values which were considered as edges. All recorded vibration measurement data for each of three axes in time domain were filtered by Median Filter in order to remove noise and spikes.

$$N_{Hz} = \frac{5000}{2 * \pi * r * 60 * 60} = 2.1897 Hz \quad (1)$$

In the formula, r is the radius of the brake drum equals to 0.101 m.

Amplitude values of all filtered vibration data were obtained by using all axes data for each time samples and it is reached to the amplitude vector as a 1-dimensional signal in time domain and transferred with FFT to frequency domain in MATLAB. The three-axis accelerometer produces simultaneous data on three axes (x , y and z) in each time slot. When this data in each time slot is considered as vector, in order to perform one-dimensional signal processing, the amplitude of the vector is taken as it requires only one value instead of three values in each time slot and calculated by Eq. (2). The amplitude or norm or length or magnitude of this vector is known as absolute value. Therefore, negative parts are not included because absolute value expressions cannot be negative. In addition, amplitudes in frequency domain are given to show which frequencies have peaks. As seen in the frequency domain representations from Fig. 3 and Fig. 4, increasing of the peak value in 213.4 Hz used by the Goertzel algorithm will also be verified in next section.

$$A_{v_i} = \sqrt{m_{x_i}^2 + m_{y_i}^2 + m_{z_i}^2} \quad (2)$$

These processes are repeated for each measurement. The collected vibration signals in time domain and frequency domain are shown in Fig. 3 and Fig. 4 with recording date. For all signals, total number of samples (N) is 8192 and sampling period (T_s) is 0.000390625 second.

III. PROPOSED APPROACH AND EVALUATIONS

The autocorrelation function is generated when one of the same copy of the signal is shifted relative to the original one and calculated by Eq. (3).

$$R_x[n] = \sum_{m=-\infty}^{\infty} x[m+n]x[m] \quad (3)$$

Autocorrelation functions for all measurements are presented in Fig. ???. According to these autocorrelation functions, the total difference method proposed with this paper was applied. According to this method, the difference between the summation of maximum peak values and minimum peak values were calculated by Eq. (4) and presented in Fig. 6.

$$D = \left(\sum_{i=1}^N A_i^{max} \right) - \left(\sum_{i=1}^N A_i^{min} \right) \quad (4)$$

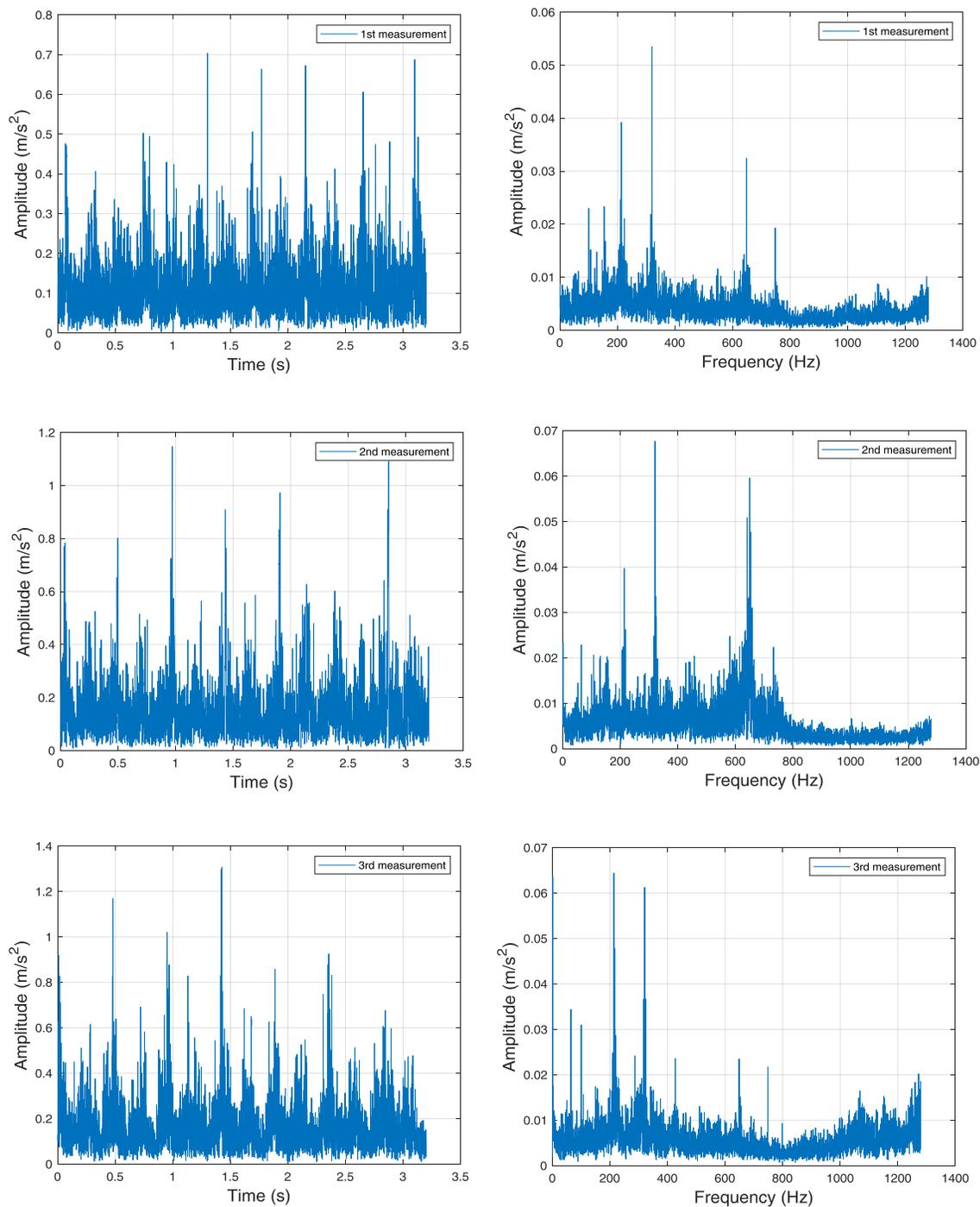


Fig. 3. Time and frequency domain representation of vibration amplitude signals of measurements of first rolling-element bearing

As can be understood from the Fig. 6, as the time passes, difference values of the vibration signals increase. The difference value before the first bearing was broken down, reached

to peak value of 5967 with the 3rd measurement. This value can be regarded as the threshold level. After the first bearing was broken down, the difference between the summation of

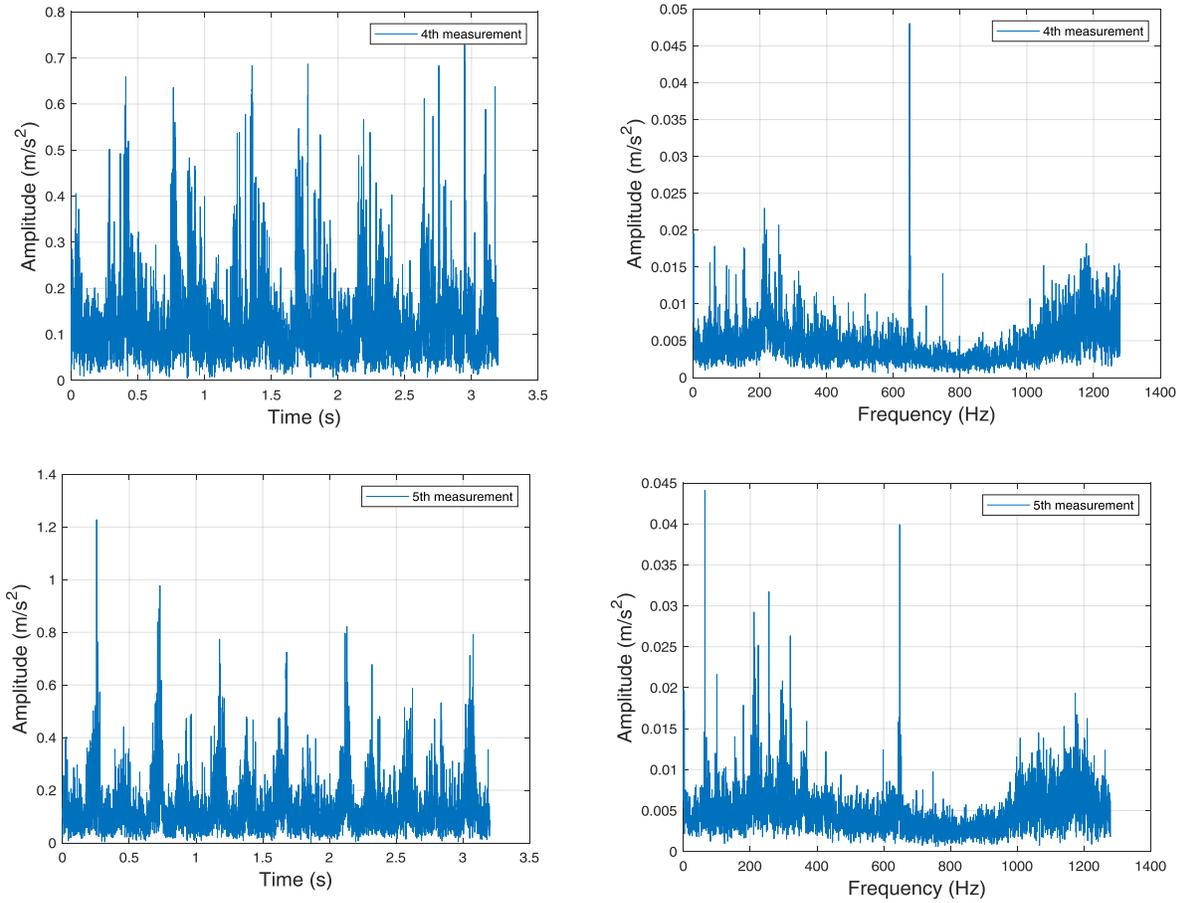


Fig. 4. Time and frequency domain representation of vibration amplitude signals of measurements of second rolling-element bearing

maximum peak values and minimum peak values for the second ball bearing increased from lower to higher value over time.

FFT algorithm is utilized in order to calculate DFT (Discrete Fourier Transform). When a few or specific DFT frequency samples are only needed, Goertzel algorithm can be used to decrease calculation complexity more whereas obtaining all the frequency samples [8]. The Goertzel algorithm is considerably superior to ensure energy efficiency against the FFT with less computational load, since less computational load causes less energy consumption [9].

Amplitude calculation of specific frequency sample with the Goertzel algorithm, which needs only signal energy without phase information, using the previous input values is calculated as

$$\begin{aligned}
 X_G^2(k) &= Q_k^2(N-1) + Q_k^2(N-2) \\
 &- 2\cos(2\pi\frac{k}{N})Q_k(N-1)Q_k(N-2)
 \end{aligned}
 \quad (5)$$

where Q_k s are intermediate values that are formed when

entering the input signals in time domain in the Goertzel algorithm.

In mechanical signal analysis literature for the rolling-element bearings, the Goertzel algorithm is utilized for the first time with this study. The Goertzel algorithm is used to obtain predetermined and specific spectral amplitude of 213.4 Hz frequency sample in this paper. 213.4 Hz frequency sample corresponds to outer raceway defect frequency harmonic and calculated by Eq. (6). As can be understood from the Fig. 7, as the time passes, spectral amplitudes of 213.4 Hz frequency samples of the vibration signals increase. The final measurement (3rd one) before the first bearing was broken down, spectral amplitude of 213.4 Hz frequency sample reached to peak value of 0.0644 m/s^2 . This value can be regarded as the threshold level. After the first bearing was broken down, spectral amplitude of 213.4 Hz frequency sample for the second ball bearing increased from lower to higher value over time.

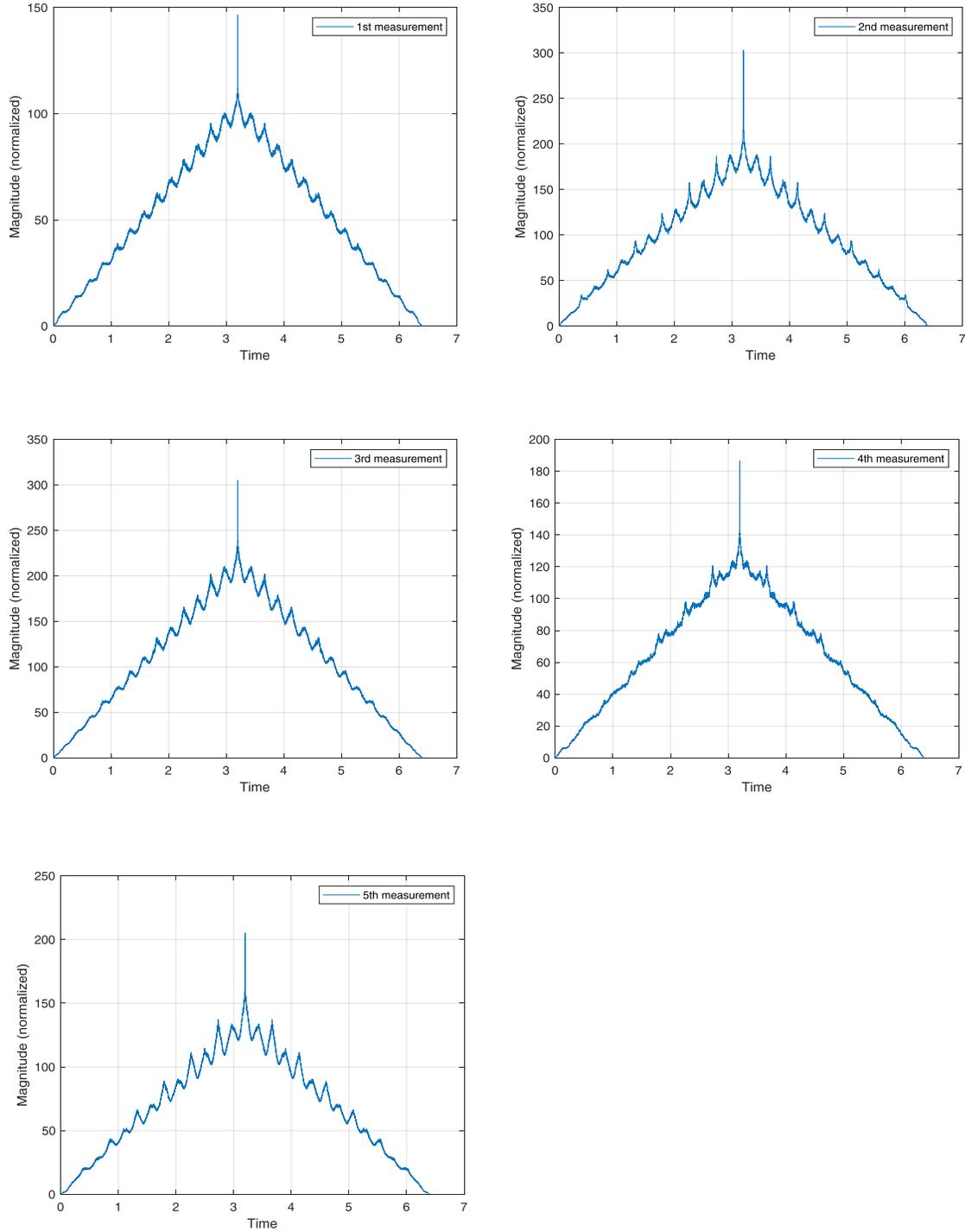


Fig. 5. Autocorrelation functions for all measurements

Outer raceway defect frequency (ORF) is calculated as 8.0771 Hz according to the formula above where N_{Hz} is the RPM of the shaft in Hz equals to 2.1897 Hz, d is the mean

$$ORF = \left(\frac{n}{2}\right)(N_{Hz})\left(1 - \frac{d}{D}\cos\alpha\right) \quad (6)$$

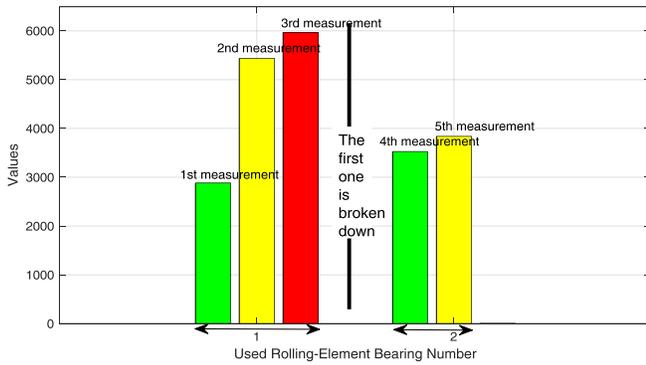


Fig. 6. The difference values according to the total difference method

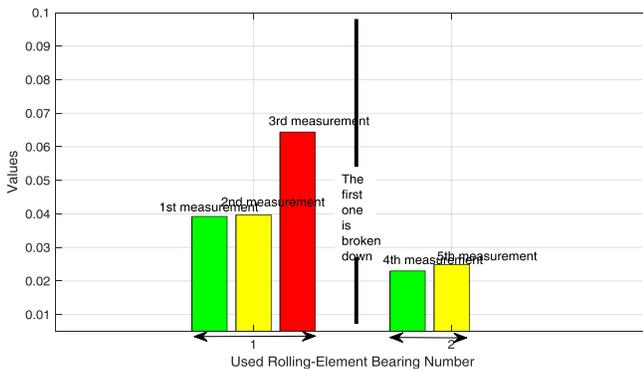


Fig. 7. Spectral amplitude values of 213.4 Hz frequency samples

diameter of the rolling elements equals to 7.40 mm, D is the pitch diameter of the bearing equals to 41 mm, n is the number of rolling elements equals to 9 and α is the contact angle equals to 0. 213.4 Hz is the 26th harmonic of ORF.

IV. CONCLUSION

Safety and efficiency of vehicle roller brake tester used at vehicle inspection stations are important for reliable inspection. Rolling element bearing used in the brake tester play important role for effective vehicle inspection tests. In this study, pre-alarming and failure analysis of rolling-element bearing of vehicle brake tester used at a vehicle inspection station was considered by taking the real time vibration signals. Real time vibration signals were subjected to autocorrelation difference and Goertzel algorithm methods after median filtering process. At this point, it will be possible to obtain an effective prediction of rolling element bearing failures in the brake testers at vehicle inspection stations.

ACKNOWLEDGMENT

The author gratefully thanks Assoc. Prof. Dr. Suat Sardemir for helping during experiments.

REFERENCES

[1] YANG, Z., U.C. MERRILD, M.T. RUNGE, G. PEDERSEN, and H. BRSTING. A study of rolling-element bearing fault diagnosis using motor's vibration and current signatures, *IFAC Proceedings Volumes*,

vol. 42, no. 8, pp. 354– 359, 2009, 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1474667016358025>

[2] MING, A., Z. QIN, W. ZHANG, and F. CHU. Spectrum auto-correlation analysis and its application to fault diagnosis of rolling element bearings, *Mechanical Systems and Signal Processing*, vol. 41, no. 1, pp. 141 – 154, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0888327013003798>

[3] BOUDIAF, A., A. DJEBALA, H. BENDJMA, A. BALASKA and A. DAHANE. A summary of vibration analysis techniques for fault detection and diagnosis in bearing, in *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*, Nov 2016, pp. 37–42.

[4] RAI, V. and A. MOHANTY, Bearing fault diagnosis using fft of intrinsic mode functions in hilberthuang transform, *Mechanical Systems and Signal Processing*, vol. 21, no. 6, pp. 2607 – 2615, 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0888327006002846>

[5] KULAC, S. and S. SARIDEMIR, Evaluation of vibration spectral values of a rolling-element bearing used in a vehicle inspection station. 3rd International Conference on Engineering and Natural Science ICENS 2017, 2017.

[6] SINGH, S. C.Q. HOWARD, C.H. HANSEN AND U.G. KOPKE. Analytical validation of an explicit finite element model of a rolling element bearing with a localised line spall, *Journal of Sound and Vibration*, vol. 416, pp. 94 – 110, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0022460X17306703>

[7] TIAN, X., J.X. GU, I. REHAB, G.M. ABDALLA, F.GU and A. BALL. A robust detector for rolling element bearing condition monitoring based on the modulation signal bispectrum and its performance evaluation against the kurtogram, *Mechanical Systems and Signal Processing*, vol. 100, pp. 167 – 187, 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0888327017304004>

[8] GOERTZEL, G. An algorithm for the evaluation of finite trigonometric series, *The American Mathematical Monthly*, vol. 65, no. 1, pp. 34–35, 1958. [Online]. Available: <http://www.jstor.org/stable/2310304>

[9] WANG, W., G. ZHIBIN, L. HUANG and Y. YAO. Spectrum sensing based on goertzel algorithm, pp. 1 – 4, 11 2008.

Modified Modular Unit Bits Sum Codes with Arbitrary Account Modules

Dmitrii V. Efanov,
DSc, Professor at "Automation, Remote
Control and Communication
on Railway Transport",
Russian University of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

Anna O. Filippochkina,
Student of the Institute of Transport
Engineering and Control Systems,
graduating sub-department «Automation,
Remote Control and Communication
in Railway Transport»,
Russian University of Transport (MIIT),
Moscow, Russia
filippochkinaanna24061998@mail.ru

Mariia V. Ivanova,
Student of the Institute of Transport
Engineering and Control Systems,
graduating sub-department «Automation,
Remote Control and Communication
in Railway Transport»,
Russian University of Transport (MIIT),
Moscow, Russia
mariaivanova27@bk.ru

Abstract—The authors analyzes the ways of unit bits sum code design with improved error detection characteristics in data vectors in comparison with known codes. It is proposed to use arbitrary account modules in the construction of a modified modular sum code, which gives different sum codes families. Installed characteristics of the modified error detection sum code with modular summing the unit bits. The detection sum codes properties as an error with different multiplicities and some errors type like a unidirectional, symmetrical and asymmetrical. The results can be used to construct systems with fault detection for the synthesis of the self-checking system and for controllable structures organization.

Keywords—unit bits sum code; modular sum code; codes with arbitrary modules account; error detection in the data vector; undetectable error; undetected errors type and multiplicity.

I. INTRODUCTION

Barred codes are widely used in the information communication and processing of organization, as well as in solving problems of technical diagnostics devices and control systems [1 – 3]. Different applications use different design and property codes. For example, some applications require correction of distortions that occur in code words, and some applications require the only detection emerged distortion. The latter problem arises, for example, in the synthesis of discrete systems, gives a fault detection feature (the synthesis of systems with self-verification and controllable structures, the organization of the working diagnosis, etc.) [4 – 6].

Codes focused on the detection of distortions have less redundancy than the codes correcting errors. This allows you to build devices with fault detection with low redundancy and the exception of the accumulation of faults. Widespread use in such problems have received various block uniform codes, and above all, various systematic codes obtained through the use of predetermined bits parity checks [7, 8] weighted codes [9] and unit bits sum codes and weighted bits sum codes [10, 11]. The designs of these codes are quite simple, in addition, well-known methods for the f coder's synthesis with self-checking structures on a different functional basis.

This paper presents the results of research in the ways development to modify the classical modular unit bits sum

codes [12 – 16] in order to improve the characteristics of their errors detection in data vectors. The features of the modified modular unit bits sum codes with arbitrary counting modules are described in detail and set their key characteristics of different types and multiplicities error detection.

II. METHODS FOR MODIFIED CODES WITH SUMMATION

The considered sum codes are modifications of the known classical modular sum codes and have improved characteristics of error detection.

Modular sum codes (denote them as $SM(m,k)$ -codes, where M is the value of the module used in the construction of the code, m and k are the number of bits in the data and check vectors, respectively) are constructed as follows [15]. The weight r of the data vector – the number of unit bits is calculated. Then the smallest nonnegative deduction of the obtained number is determined by the preset module $M \in \{2; 3; \dots; m+1\}$ – the number $W = r(\text{mod } M)$ is obtained. This number is represented in binary form and is written to the bits of the check vector. The number of control bits in the codewords $SM(m,k)$ -codes is determined by the value $k = \lceil \log_2 M \rceil$ (the entry $\lceil \dots \rceil$ denotes an integer at the top of the calculated value). As shown in [17], $SM(m,k)$ -codes for $M \in \{2^1; 2^2; \dots; 2^{\lceil \log_2(m+1) \rceil}\}$, has better error detection characteristics among all codes with the same number of check bits. Also, for these codes are most simply synthesized coders with self-checking structures. However, in some applications, $SM(m,k)$ -codes with arbitrary counting modules can be effective [15]. The authors consider the modified modular unit bits sum codes, which are built from the "basic" modular sum codes. When constructing these codes, the modified weight of the data vector is determined by the formula:

$$W = r(\text{mod } M) + \alpha M, \quad (1)$$

In (1) α – a special correction index calculated as the sum modulo two (XOR) pre-selected data bits. The first term in (1) defines, in fact, the "basic" $SM(m,k)$ -code and "basic" properties of the modified code under construction. The second term allows you to improve the properties of the "base" code.

Modified modular unit bits sum codes are denoted as $RSM(m,k)$ -codes, separately defining the formula for calculating the correction coefficient α .

The considered method of constructing a modified sum code is known in world literature. For example, in [11] it is proposed to use this method of modification, except that the smallest non-negative deduction is determined for the total value of the weight coefficients of the bits or transitions between the bits. In this case, the correction factor is proposed to be calculated by the formula: $\alpha = x_m \oplus x_{m-1} \oplus \dots \oplus x_{k+1}$ (low order k bits do not participate in the determination of the correction index). In [18] the modified code with summation is considered, for which (1) is directly used, the module is determined by the value $M = 2^{\lceil \log_2(m+1) \rceil - 1}$, and the coefficient α by the same formula: $\alpha = x_m \oplus x_{m-1} \oplus \dots \oplus x_{k+1}$. In [19] the idea of building, this code is common to the use of modules $M = \{2^1, 2^2, \dots, 2^{\lceil \log_2(m+1) \rceil - 2}\}$. The correction index α is proposed to be calculated by the formula: $\alpha = x_m \oplus x_{m-1} \oplus \dots \oplus x_p$, $p = \lceil \log_2(m+1) \rceil + 2$. In article [20] are constructed families of $RSM(m,k)$ -codes with different values of α and modules $M \in \{2^1, 2^2, \dots, 2^{\lceil \log_2(m+1) \rceil - 1}\}$. In all of these studies are presented only ways of code construction and some common characteristics of error detection by multiplicities.

In many sum codes, applications are important not only features of detecting errors of different multiplicities but also types [21 – 25]. From this point of view, it is important to study the characteristics of unidirectional, symmetrical and asymmetrical errors detection [26]. The unidirectional error occurs when only zero or only one bits of the code word or data vector are distorted. Symmetrical error is associated with the same number of zero and one bits distortions. The asymmetrical error occurs in the case of an unequal number of distortions of zero and one bits. Different sum codes have different characteristics of error detection in data vectors both by multiplicities and by their types.

Next, we consider $RSM(m,k)$ -codes with random modules of calculation $M \in \{2, 3, 4, \dots, 2^{\lceil \log_2(m+1) \rceil - 1}\}$ and methods for calculating the correction index α (Fig. 1).

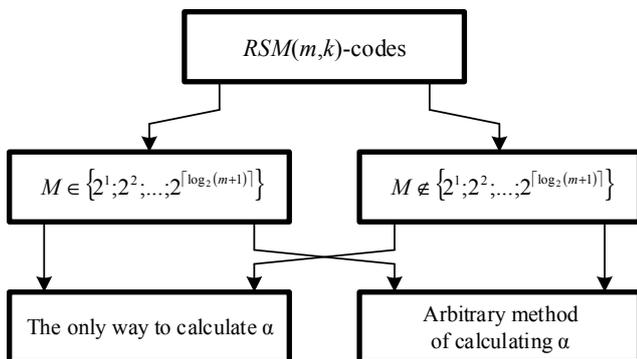


Fig. 1. Classification of $RSM(m,k)$ -codes.

III. MODIFIED MODULAR CODES WITH ARBITRARY ACCOUNT MODULES

When you modify the $SM(m,k)$ -code in the $RSM(m,k)$ -code, as noted above, the basic characteristics of error detection are preserved. This is due to the method of calculat-

ing the modified weight, which is convenient to explain by the example of the distribution of all data vectors between all check vectors. For example, this distribution is given in Tabl. 1 and Tabl. 2 for $S3(4,2)$ and $RS3(4,3)$ -codes (the correction index is determined by the formula $\alpha = x_4 \oplus x_3$, the numbering of bits – from right to left).

TABLE I. $S3(4,2)$ - CODE

Check vectors		
00	01	10
Reference groups		
0	1	2
Data vectors		
0000	0001	0011
0111	0010	0101
1011	0100	0110
1101	1000	1001
1110	1111	1010
		1100

TABLE II. $RS3(4,3)$ -CODE, $\alpha = x_4 \oplus x_3$

Check vectors							
000	001	010	011	100	101	110	111
Reference groups							
0	1	2	3	4	5	6	7
Data vectors							
0000	0001	0011	0111	0100	0101		
1101	0010	1100	1011	1000	0110		
1110	1111				1001		
					1010		

The source Table of the distribution of data of vectors between the check vectors (in reference groups) is expanding due to the "shear" part of the data vectors, for which $\alpha = x_4 \oplus x_3$, into reference groups with greater numbers. This shift is carried out by the value $s = M$. With this modification of the $SM(m,k)$ -code, data vectors are not added to the existing reference groups, and some of the available vectors occupy empty groups. Since the error will not be detected only if it translates any data vector of one reference group into a data vector of the same reference group, the number of undetectable errors decreases with a decrease in the number of data vectors in the group. However, the basic properties of the code are preserved.

The above reasoning is followed by important findings (given without proof).

Theorem 1. Any $RSM(m,k)$ -code, regardless of the values of the modulus M and the procedure for the calculation of the correction index α detects a greater number of symmetrical errors in the data vectors than the corresponding modular code with summation, and any unidirectional errors, with the exception of a percentage of the unidirectional errors with multiplicity:

$$d = jM, \quad j=1,2,\dots,\left\lfloor \frac{m}{M} \right\rfloor \quad (2)$$

Theorem 2. Any $RSM(m,k)$ -code, regardless of the value of the module M and the rules for calculating the correction index α , detects any asymmetrical errors in the data vectors, except for a certain proportion of asymmetrical errors with multiplicity:

$$d = M + 2j, \quad j=1,2,\dots,\left\lfloor \frac{m-M}{2} \right\rfloor \quad (3)$$

These properties are characteristic also for the classical $SM(m,k)$ -codes. Both codes belong to d_0, d_α - $UAED(m,k)$ -codes (unidirectional and asymmetrical error-detection codes) [27], where $d_0=M$, a $d_\alpha=M+2$ – are minimal multiplicities of undetectable unidirectional and asymmetrical errors, respectively.

Note that all 2^k reference groups can be filled with at least one data vector only if $M \in \{2^1; 2^2; \dots; 2^{\lceil \log_2(m+1) \rceil - 1}\}$. For other codes, part of the reference groups will always be empty. This complicates the procedure of coders synthesis with self-checking structures for codes with $M \notin \{2^1; 2^2; \dots; 2^{\lceil \log_2(m+1) \rceil - 1}\}$.

Since some groups will always be empty at $M \notin \{2^1; 2^2; \dots; 2^{\lceil \log_2(m+1) \rceil - 1}\}$, some alternative ways of constructing $RSM(m,k)$ -codes can be proposed. "Shift" of data vectors can be carried out not only by the value $s = M$. Other modified modular sum codes can be constructed using the formula:

$$W = r(\text{mod } M) + \alpha s, \quad (4)$$

where $s \in \{M; M+1; \dots; 2^{\lceil \log_2 M \rceil + 1} - M\}$.

For example, for the above $S3(4,2)$ -code can be modified with shifts of $s \in \{3; 4; 2^{\lceil \log_2 3 \rceil + 1} - 3 = 5\}$. If $s = 4$ reference groups No. 3 and No. 7 will always be empty, and when $s = 5$ – No. 3 and No. 4.

Various methods of constructing $RSM(m,k)$ -codes with values $M \notin \{2^1; 2^2; \dots; 2^{\lceil \log_2(m+1) \rceil - 1}\}$ can be used under restrictions on the formation of check vectors, necessary for a complete check of the structure of the encoder.

TABLE III. $RSM(M,K)$ -CODES WITH A GIVEN NUMBER OF CHECK BITS

Number of check bits	Codes
1	–
2	$RS2$
3	$RS3, RS4$
4	$RS5 \dots RS8$
5	$RS9 \dots RS16$
...	...
k	$RS(2^{k-1}+1) \dots RS(2^k-1)$

Table 3 provides a classification of the codes in question by the number of check bits. Note also that to reduce the number of check bits when choosing the module $M \notin \{2^1; 2^2; \dots; 2^{\lceil \log_2(m+1) \rceil - 1}\}$ you can additionally determine the smallest non-negative deduction of the number W

$M^* = 2^{\lceil \log_2(m+1) \rceil - 1}$. However, this method will give the noise-immune code only at a certain initial "shift" (at a certain value of the number s).

IV. PROPERTIES OF MODULAR CODES WITH ARBITRARY ACCOUNT MODULES

A. Codes with an arbitrary number of data bits in the amount of correction index

Analysis of the table form for specifying the $RSM(m,k)$ -code (see Tabl. 2) allows you to set the characteristics of the detection of errors in data vectors by different types and multiplicities. For example, let us analyze possible undetectable errors that are obtained by erroneous transitions of data vectors into each other within the reference group No. 0 of Tabl. 2. There are three vectors in this control group: $\langle 0000 \rangle$, $\langle 1101 \rangle$ и $\langle 1110 \rangle$. The transitions between the vectors $\langle 0000 \rangle$ and $\langle 1101 \rangle$ occur at three times the unidirectional error, transitions between the vectors $\langle 0000 \rangle$ and $\langle 1110 \rangle$ – also with the three unidirectional error, and the transitions between the vectors $\langle 1101 \rangle$ and $\langle 1110 \rangle$ – the two symmetrical error. All in all, this reference group gives 6 undetectable errors – 4 triple-bit unidirectional error and two double-bit symmetrical errors.

Automation analysis of the tabular form of the task $RSM(m,k)$ -code allowed the authors to study the regularities that appear in different ways of design codes with different values of modules, lengths of data vectors and the number of data bits in the sum of the correction index.

As an example in Tabl. 4 is given the number of undetectable errors in data vectors $RS5(8,3)$ -codes with different ways of calculating the correction index.

Because each bit of data vector takes only two possible values (0 or 1), and analyzing the tabular form specifying code addresses all possible 2^m data vectors, then the error detection features $RSM(m,k)$ -codes will be determined not by what data bits are included in the amount of the correction index, but only their number $b(\alpha)$. For reason, all the codes with the same value $b(\alpha)$ form a single subclass $RSM(m,k)$ -codes for the specified values m and M .

Analysis of characteristic tables for $RSM(m,k)$ -codes with different values of m and M allowed establishing the features of error detection in data vectors by these codes. These features are inherent in all $RSM(m,k)$ -codes with arbitrary account modules:

1. $RSM(m,k)$ -codes with even values m detect any errors with even multiplicities.
2. $RSM(m,k)$ -codes with odd values of m do undetected some fraction of errors with odd multiplicities. These codes detect any errors with odd multiplicities $d \leq M-2$.
3. $RSM(m,k)$ -codes $b(\alpha) = j$ and $b(\alpha) = m - j$ have the same characteristics of error detection in data vectors.
4. The total number of undetectable errors for $RSM(m,k)$ -codes with values $M=2$ и $M=4$ for a particular value m is constant regardless of the value $b(\alpha)$.
5. For $RSM(m,k)$ -codes with any values M , except $M=2$ and $M=4$, and $M=4$, for a particular value m

with increasing number $b(\alpha)$ the total number of undetectable errors decreases.

6. $RSM(m,k)$ -codes with values $M=2$ и $M=4$ for a particular value m are constant regardless of the value $b(\alpha)$.
7. For $RSM(m,k)$ -codes with any values of M regardless of the length of the data vector with increasing number $b(\alpha)$ decreases the proportion of undetectable twofold errors of the total number.
8. For $RSM(m,k)$ -codes with values $M \geq 3$ the fractions of undetectable double-bit errors from the total number of double errors for the same values m and $b(\alpha)$ are the same (see Tabl. 5).
9. For $RSM(m,k)$ -codes with values $M \geq 5$ are the same fractions of undetectable fourfold errors of the total number of fourfold errors for the same values m and $b(\alpha)$.
10. For any $RSM(m,k)$ -code with a specific value m with an increase in the number $b(\alpha)$ there is a decrease in the values of the fractions of undetectable symmetrical errors from their total number.
11. The proportion of unidirectional errors from their total number for any $RSM(m,k)$ -code increases slightly with increasing $b(\alpha)$, and then decreases; the greatest number of unidirectional undetectable error occurs in the codes with $b(\alpha)=1$.
12. The fraction of asymmetrical errors from the total number of errors for $b(\alpha)=1$ for any $RSM(m,k)$ -code is minimal, with increasing $b(\alpha)$ decreases slightly, and then increases slightly.
13. With increasing length of the data vector for each $RSM(m,k)$ -code there is a gradual increase in the fraction of unidirectional undetectable errors of the total number, reaching a maximum at a certain value m .
14. The values of the shares of undetectable $RSM(m,k)$ -codes of unidirectional and asymmetrical errors from the total number of errors decrease with the increase of the module value, and the share of symmetrical errors from their total number fluctuates around 50%.
15. The best characteristics for error detection in data vectors have such $RSM(m,k)$ -codes for which

$$b(\alpha) = \left\lfloor \frac{m}{2} \right\rfloor.$$

B. Codes with the smallest total number of undetectable errors

Studies show that among all $RSM(m,k)$ -codes can be allocated codes for which the minimum total number of undetectable errors in the data vectors for a given value m . They are constructed according to the rules given earlier considering that in the sum of correction index are used

$b(\alpha) = \left\lfloor \frac{m}{2} \right\rfloor$ data bits. This class of codes is promising for solving problems of technical diagnostics of discrete systems. Analysis of characteristic tables for $RSM(m,k)$ -codes, which $b(\alpha) = \left\lfloor \frac{m}{2} \right\rfloor$, allowed to establish the following key properties of this class of codes:

1. With the increase in the value of M for codes with the same number of control digits, the proportion of undetectable errors from the total number of them for the same values of m decreases, and the efficiency coefficient, respectively, increases.
2. As the value of M for codes with the same number of check bits increases, the proportion of undetectable unidirectional and asymmetrical errors from the total number of errors of the respective type for the same values of m decreases.
3. The fraction of symmetrical undetectable errors from the total number of errors does not depend on the value of M and is determined only by the length of the data vector.
4. The fraction of double-bit undetectable errors from the total number of double-bit errors also does not depend on the value of M and is determined only by the length of the data vector.

Interesting are the characteristics of error detection $RSM(m,k)$ -codes for families of codes with the same number of check bits. Let's consider the main ones on the example of codes with $k=3$.

In Figures 2 – 7 show graphs of such indicators as:

– $\gamma_{m,k}$ – the fraction of undetectable errors in data vectors of codes from the total number of errors in data vectors; $\xi_{m,k}$ – the efficiency coefficient of using check bits, calculated by the formula: $\xi_{m,k} = \frac{N_{m,k}^{\min}}{N_{m,k}}$, where

$N_{m,k}^{\min} = 2^m(2^{m-k} - 1)$ – the minimal total number of undetectable errors for specific values of m and k , $N_{m,k}$ – the number of undetectable errors in code [20];

– $\nu_{m,k}$, $\sigma_{m,k}$, $\alpha_{m,k}$ – the fraction of undetectable unidirectional, symmetrical and asymmetrical errors in data vectors of codes from the total number of errors of this kind in data vectors;

– $\beta_{m,2}$ – the fraction of double-bit undetectable errors in data vectors of codes from the total number of double-bit errors in data vectors.

From the graphs in Fig. 2 and Fig. 3 it follows that among all $RSM(m,k)$ -codes with the same number of control bits, the best error detection properties are those codes for which the number $M = 2^k$ is selected as a module. The number of undetectable errors is reduced with increasing values of the modulus and the approximation to this value. Similar regularities are manifested when considering the characteristics of detection of unidirectional and asymmetrical errors by $RSM(m,k)$ -codes: as the value of M increases, the fraction of undetectable unidirectional and asymmetrical errors from the total number of errors of these kinds decreases, and their maximum is shifted towards increasing the length of the data vector. The fraction of undetectable symmetrical errors for these codes and the proportion of double-bit undetectable errors from the total number of them are constant regardless of the value of M .

TABLE IV. CHARACTERISTIC TABLE FOR $RS(8,3)$ - CODES WITH DIFFERENT VALUES OF $B(A)$

$b(\alpha)$	Number of undetectable errors with multiplicities d								Total number of undetectable errors
	1	2	3	4	5	6	7	8	
1	0 0/0/0	2688 0/2688/0	0 0/0/0	3360 0/3360/0	336 336/0/0	560 0/560/0	28 0/0/28	0 0/0/0	6972 336/6608/28
2	0 0/0/0	2048 0/2048/0	0 0/0/0	2880 0/2880/0	416 416/0/0	1280 0/1280/0	168 0/0/168	70 0/70/0	6862 416/6278/168
3	0 0/0/0	1664 0/1664/0	0 0/0/0	3360 0/3360/0	496 496/0/0	1200 0/1200/0	84 0/0/84	0 0/0/0	6804 496/6224/84
4	0 0/0/0	1536 0/1536/0	0 0/0/0	3648 0/3648/0	448 448/0/0	960 0/960/0	112 0/0/112	70 0/70/0	6774 448/6214/112
5	0 0/0/0	1664 0/1664/0	0 0/0/0	3360 0/3360/0	496 496/0/0	1200 0/1200/0	84 0/0/84	0 0/0/0	6804 496/6224/84
6	0 0/0/0	2048 0/2048/0	0 0/0/0	2880 0/2880/0	416 416/0/0	1280 0/1280/0	168 0/0/168	70 0/70/0	6862 416/6278/168
7	0 0/0/0	2688 0/2688/0	0 0/0/0	3360 0/3360/0	336 336/0/0	560 0/560/0	28 0/0/28	0 0/0/0	6972 336/6608/28

Note. In each cell of the table, the total number of errors is written on the top, and the number of unidirectional, symmetrical and asymmetrical errors is written on the bottom through the slash, respectively.

TABLE V. VALUES OF $B_{m,2}$ FOR $RSM(M,K)$ -CODES WITH VALUES $M \geq 3$

m	$b(\alpha)$									
	1	2	3	4	5	6	7	8	9	10
4	25	16.667								
5	30	20								
6	33.333	23.333	20							
7	35.714	26.19	21.429							
8	37.5	28.571	23.214	21.429						
9	38.889	30.556	25	22.222						
10	40	32.222	26.667	23.333	22.222					
11	40.909	33.636	28.182	24.545	22.727					
12	41.667	34.848	29.545	25.758	23.485	22.727				
13	42.308	35.897	30.769	26.923	24.359	23.077				
14	42.857	36.813	31.868	28.022	25.275	23.626	23.077			
15	43.333	37.619	32.857	29.048	26.19	24.286	23.333			
16	43.75	38.333	33.75	30	27.083	25	23.75	23.333		
17	44.118	38.971	34.559	30.882	27.941	25.735	24.265	23.529		
18	44.444	39.542	35.294	31.699	28.758	26.471	24.837	23.856	23.529	
19	44.737	40.058	35.965	32.456	29.532	27.193	25.439	24.269	23.684	
20	45	40.526	36.579	33.158	30.263	27.895	26.053	24.737	23.947	23.684

Note. The presented values for $RSM(m,k)$ -codes with the number of check bits $k=3$ can be compared with similar values for classical $SM(m,k)$ -codes, for which $\beta_{m,2}=50\%$ regardless of the length of the data vector [17].

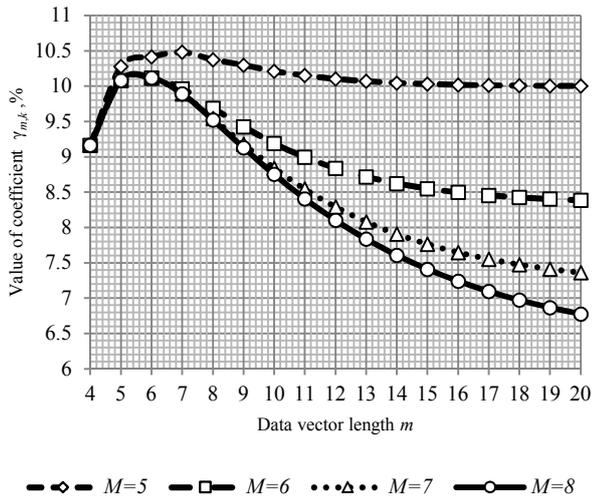


Fig. 2. Dependence of the value of $\gamma_{m,k}$ for $RSM(m,k)$ -codes with the number of check bits $k=3$.

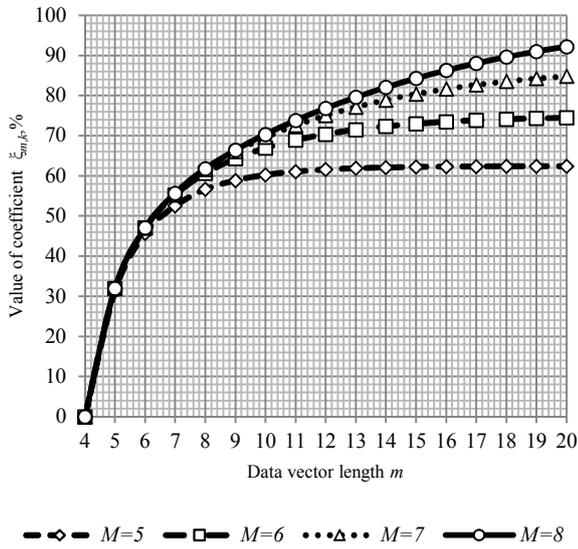


Fig. 3. Dependence of the value of $\xi_{m,k}$ for $RSM(m,k)$ -codes with the number of check bits $k=3$.

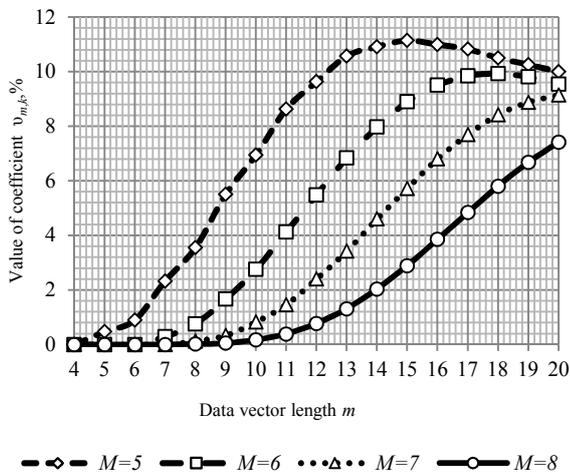


Fig. 4. Dependence of the value $v_{m,k}$ for $RSM(m,k)$ -codes with the number of check bits $k=3$.

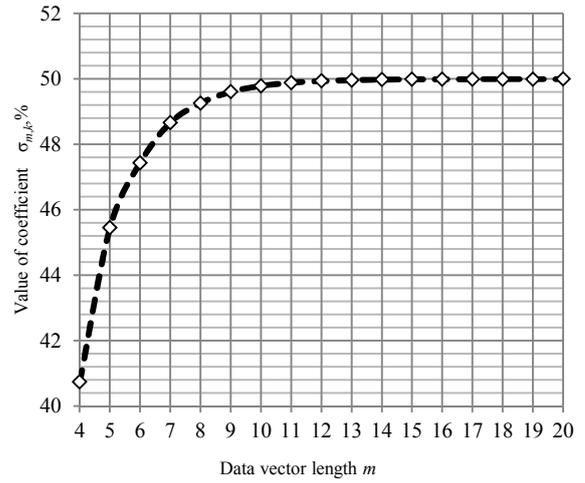


Fig. 5. Dependence of the value $\sigma_{m,k}$ for $RSM(m,k)$ -codes with the number of check bits $k=3$.

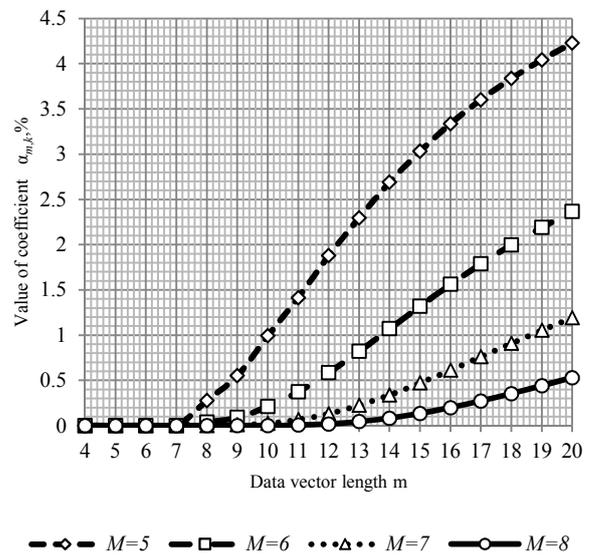


Fig. 6. Dependence of the value $\alpha_{m,k}$ for $RSM(m,k)$ -codes with the number of check bits $k=3$.

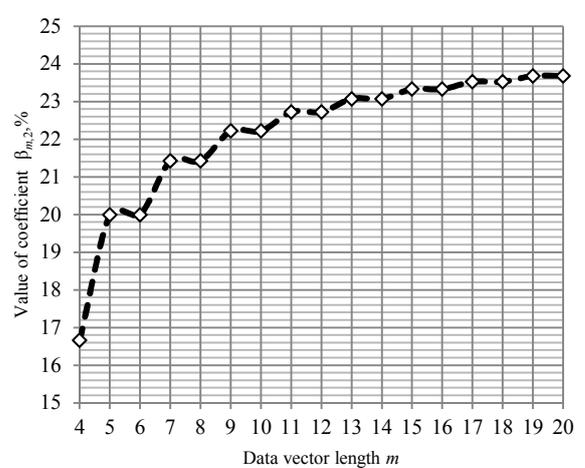


Fig. 7. Dependence of the value $\beta_{m,2}$ for $RSM(m,k)$ -codes with the number of check bits $k=3$.

V. CONCLUSION

This study for the first time addresses the issue of a comprehensive study of the characteristics of error detection of different types and with different multiplicities in the data vectors of $RSM(m,k)$ -codes with arbitrary account modules.

Assessing the results, we can draw the following conclusions. First, the class of $RSM(m,k)$ -codes is significantly expanded by choosing degrees not only from the set $M \in \{2^1; 2^2; \dots; 2^{\lceil \log_2(m+1) \rceil - 1}\}$. Second, the number of $RSM(m,k)$ -codes with different characteristics of the error detection is not so great and is only determined by three important factors: the values of the numbers m , M и $b(\alpha)$. At the third, all $RSM(m,k)$ -codes form the class d_v, d_α -UAED(m,k)-codes, where $d_v=M$ and $d_\alpha=M+2$. These key regularities inherent in modularly modified unit bits sum codes allow determining the standard of their applicability in solving the problems of discrete systems synthesis with fault detection and their technical diagnostics.

REFERENCES

- [1] E. Fujiwara "Code Design for Dependable Systems: Theory and Practical Applications", John Wiley & Sons, 2006, 720 p.
- [2] M. Gössel, V. Ocheretny, E. Sogomonyan, and D. Marienfeld "New Methods of Concurrent Checking: Edition 1", Dordrecht: Springer Science+Business Media B.V., 2008, 184 p.
- [3] W.E. Ryan, and S. Lin "Channel Codes: Classical and Modern", Cambridge University Press, 2009, 708 p.
- [4] S. Mitra, and E.J. McCluskey "Which Concurrent Error Detection Scheme to Choose?", Proceedings of International Test Conference, 2000, USA, Atlantic City, NJ, 03-05 October 2000, pp. 985-994, doi: 10.1109/TEST.2000.894311.
- [5] P.K. Lala "Self-Checking and Fault-Tolerant Digital Design", San Francisco: Morgan Kaufmann Publishers, 2001, 216 p.
- [6] V. Hahanov "Cyber-Physical Computing for IoT-driven Services", New York, Springer International Publishing AG, 2018, 279 p.
- [7] F.J. MacWilliams, and N.J.A. Sloane "The Theory of Error-Correcting Codes", Amsterdam: North-Holland, 1977, 785 p.
- [8] A. Stempkovskiy, D. Telpukhov, S. Gurov, T. Zhukova, and A. Demeneva "R-Code for Concurrent Error Detection and Correction in the Logic Circuits", 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 29 January – 1 February 2018, Moscow, Russia, pp. 1430-1433, doi: 10.1109/EIConRus.2018.8317365.
- [9] S.J. Piestrak "Design of Self-Testing Checkers for Unidirectional Error Detecting Codes", Wrocław: Ofiyna Wydawnicza Politechniki Wrocławskiej, 1995, 111 p.
- [10] D. Das, N.A. Touba, M. Securing, and M. Gossel "Low-Cost Concurrent Error Detection Based on Modulo Weight-Based Codes", Proceedings of the IEEE 6th International On-Line Testing Workshop (IOLTW), Spain, Palma de Mallorca, July 3-5, 2000, pp. 171-176, doi: 10.1109/OLT.2000.856633.
- [11] V.B. Mekhov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "Checking of Combinational Circuits Basing on Modification Sum Codes", Automation and Remote Control, 2008, vol. 69, issue 8, pp. 1411-1422, doi: 10.1134/S0005117908080134.
- [12] B. Bose, and D.J. Lin "Systematic Unidirectional Error-Detection Codes", IEEE Transaction on Computers, Vol. C-34, November 1985, pp. 1026-1032.
- [13] D. J. Lin, and B. Bose "Theory and Design of error Correcting and d(d>t)-unidirectional Error Detecting (t-EC d-UED) Codes", IEEE Transaction on Computers, 1988, Vol. 37, issue 4, Pp. 433-439, doi: 10.1109/12.2187.
- [14] D. Das, and N. A. Touba "Synthesis of Circuits with Low-Cost Concurrent Error Detection Based on Bose-Lin Codes", Journal of Electronic Testing: Theory and Applications, 1999, vol. 15, issue 1-2, pp. 145-155, doi: 10.1023/A:1008344603814.
- [15] V. Sapozhnikov, VI. Sapozhnikov, and D. Efanov "Modular Sum Code in Building Testable Discrete Systems", Proceedings of 13th IEEE East-West Design & Test Symposium (EWDTS'2015), Batumi, Georgia, September 26-29, 2015, pp. 181-187, doi: 10.1109/EWDTS.2015.7493133.
- [16] P. N. Bibilo, and D. A. Gorodetskii "Automated Design of Modular Arithmetic Devices: Might CAD Replace an Engineer", Automatic Control and Computer Sciences, 2009, Vol. 43, Issue 2, pp. 63-73, doi: 10.3103/S0146411609020023.
- [17] D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "Application of Modular Summation Codes to Concurrent Error Detection Systems for Combinational Boolean Circuits", Automation and Remote Control, 2015, Vol. 76, Issue 10, Pp. 1834-1848, doi: 10.1134/S0005117915100112.
- [18] A.A. Blyudov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "A Modified Summation Code for Organizing Control of Combinatorial Circuits", Automation and Remote Control, 2012, Vol. 73, Issue 1, pp. 153-160, doi: 10.1134/S0005117912010122.
- [19] A.A. Blyudov, D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "Summation Codes for Organization of Control of Combinational Circuits", Automation and Remote Control, 2013, Vol. 74, Issue 6, pp. 1020-1028, doi: 10.1134/S0005117913060118.
- [20] A.A. Blyudov, D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "On Codes with Summation of Unit Bits in Concurrent Error Detection Systems", Automation and Remote Control, 2014, Vol. 75, Issue 8, Pp. 1460-1470, doi: 10.1134/S0005117914080098.
- [21] E.S. Sogomonyan, and M. Gössel "Design of Self-Testing and On-Line Fault Detection Combinational Circuits with Weakly Independent Outputs", Journal of Electronic Testing: Theory and Applications, 1993, vol. 4, issue 4, pp. 267-281, doi: 10.1007/BF00971975.
- [22] F.Y. Busaba, and P.K. Lala "Self-Checking Combinational Circuit Design for Single and Unidirectional Multibit Errors", Journal of Electronic Testing: Theory and Applications, 1994, vol. 5, issue 1, pp. 19-28, doi: 10.1007/BF00971960.
- [23] A. Morosow, V.V. Sapozhnikov, VI.V. Sapozhnikov, and M. Goessel "Self-Checking Combinational Circuits with Unidirectionally Independent Outputs", VLSI Design, 1998, vol. 5, issue 4, pp. 333-345, doi: 10.1155/1998/20389.
- [24] A.Yu. Matrosova, I. Levin, and S.A. Ostanin "Self-Checking Synchronous FSM Network Design with Low Overhead", VLSI Design, 2000, vol. 11, issue 1, pp. 47-58, doi: 10.1155/2000/46578.
- [25] D.V. Efanov, V.V. Sapozhnikov, and VI.V. Sapozhnikov "Synthesis of Self-Checking Combinational Devices Based on Allocating Special Groups of Outputs", Automation and Remote Control, 2018, Vol. 79, issue 9, pp. 1609-1620, doi: 10.1134/S0005117918090060.
- [26] V.V. Sapozhnikov, VI.V. Sapozhnikov, and D.V. Efanov "Errors Classification in Information Vectors of Systematic Codes" (in Russ.), Journal of Instrument Engineering, 2015, Vol. 58, Issue 5, pp. 333-343, doi: 10.17586/0021-3454-2015-58-5-333-343.
- [27] D. Efanov, V. Sapozhnikov, and VI. Sapozhnikov "The Use of Codes with Fixed Multiplicities of Detected Unidirectional and Asymmetrical Errors in the Process of Organizing Combinational Circuit Testing", Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS'2018), Kazan, Russia, September 14-17, 2018, pp. 114-122, doi: 10.1109/EWDTS.2018.8524768.

Systems for Reflectometry Analysis of Defects in Metal Structures of Transport Mobile Objects

Julia V. Alevetdinova,
Assistant, Department
of "Track, constructed machines and robotics systems",
Russian University of Transport (MIIT)
Moscow, Russia
yulialevetdinova@gmail.com

Abstract—While solving the problem of signal detection, it's necessary to lead it to digital form. The article discusses fatigue damage detection systems with linear correction of channel gain, as well as an algorithm for detecting defects from the library of algorithms. The rate of the change assessment in the metalwork relief becomes one of the most important factors since it will affect the critical state scanning. The analysis of the assessment results should occur with high accuracy since it can provide objective data on the degree of fatigue damage to the metal structure under study and evaluate its resource before failure. In the zone of expected crack development, control platform is prepared, which represent a surface area treated with high purity. Noises that are possible at the diagnostic platform can affect the quality of diagnostics, therefore, it is very important to prepare the surface for the study. The developed system of technical diagnostics of the metal structure state is based on the reflectometric sensor work, which implements the optical diagnostic method. The used reflectometric sensor has a compact size, with constant quality control of the metal-structure surface. It allows you to study the required area and conduct research on its condition. The sensor is necessary because it isn't uncommon for a defect in the surface of the metal structure, which could have been found earlier, and would not lead to negative consequences.

Keywords—*correction, algorithm, 3D model, indicatrix of scattering, surface defects, measurement error, optical method for monitoring surface conditions.*

I. INTRODUCTION

With the emergence of the need to convert an analog signal to digital, the method is appeared to carry out this procedure and device. In the process of converting an analog signal to digital, there is a high probability of noise, which negatively affects the study of metal structures. Some types of signal-to-noise ratios can be used while choosing a diagnostic method, which is discussed in more detail later in the article.

The purpose of the following article is to delve into the existing systems of technical diagnostics most suitable for detecting metal structure surface defects, to study in detail an integrated approach to the diagnosis of a metal product, for analysis. One of the main tasks is to improve the accuracy of diagnosis. Since the simultaneous use of various methods may reduce the quality of research, which is unacceptable while controlling the surface. It's also a task to develop a system for monitoring the state of the metal structure using the reflectometric diagnostic method.

It's necessary to examine in detail the existing methods for controlling the surface of the metalwork.

Adequately select a diagnostic system that is optimally suitable for conducting research of the product.

In the research of metal in the dynamics noted the manifestation of its critical state. Moreover, it should be noted that each stage of deterioration corresponds to its own changes in the surface relief.

The basis of various systems for diagnosing metal for defects is the process of non-destructive testing. Defects such as discontinuities are a consequence of the materials structure imperfection. It arises at different stages of the technological process and during operation.

The physical processes of various fields interaction, radiation or chemicals with objects of control, implement the diagnostic method itself. According to such differences, nine main types are distinguished: magnetic method, electric method, eddy current method, radio - wave method, thermal method, optical method, radiation method, acoustic method, penetrating substances.

To implement the work of these types, different control methods are used, which are classified according to the nature of the physical fields interact with the object under control, according to the primary informative parameter and the method of obtaining information.

For engineering products with a large variety of materials used in them, with different physical and mechanical properties, methods and technological processes of their manufacture, it is necessary to use a set of complementary methods and means of metal structures non-destructive testing.

Solving the issue of automation is important in order to receive information about the quality of controlled objects in electronic form. In automated NC tools, all processes are performed automatically without operator participation.

Of particular difficulty are scanning systems that are used in mechanical engineering where disassembly of structures is impossible and the approach to controlled surfaces of complex configuration is difficult. The scanning process should maintain a constant gap between the transducer, the field source and the item being monitored. The movement of the transducer and the controlled product relative to each other can be translational, rotational, complex reciprocating, etc. Scanning systems require high precision manufacturing. The mass production of industrial robots and manipulators made it possible to create on this basis various technological complexes of NC. The basis of their creation is a combination of commercially available NC devices that have access to a

computer. Industrial robots that perform the functions of moving the sensor of the device relative to the object.

There are several defect detection systems, the principle of operation, which is based on the reflectometric method for monitoring the state of the metal surface. The systems make possible to simplify the monitoring automation process. That makes possible to reduce the time spent on diagnostics, and to increase the objectivity of the evaluation of the residual life of the metal structure.

Diagnostic systems technologies provide an electronic presentation of all the data and documents used to describe a product or how it is produced and operated, for information support of various procedures used throughout the product life cycle.

The advantages of an integrated approach are comprehensive studies that can be carried out in a shorter time. In the process, the diagnostic system uses several non-destructive testing methods at once, which are examined from various points. There is also the possibility of operating an intelligent diagnostic system. This system independently decides on the use of a particular diagnostic tool.

The disadvantages of complex control include the complexity of this system implementation and the necessary work high accuracy, which the operator should follow. With the exclusion of the human factor, it is also necessary to exclude the failure factor of the system. Thus, only a comprehensive analysis will help by giving the most adequate assessment of the damage stage to the metal structure.

Conducting a different type of diagnosing the metal structures state uses already existing technical condition monitoring systems. Such systems are applicable strictly to a specific type of metalwork, and perform strictly designated control functions. In contrast to existing control systems complexes, the proposed system type for diagnosing a technical condition is universal. The system is applicable to control the various types state of metal structures used in mechanical engineering. The system is used to monitor the status of various types of metal structures used in mechanical engineering.

1. The main purpose of the multifunctional KTSM-02 complex equipment is to control rolling stock parameters that are tied to specific axes, as well as coordinating the subsystems work connected to it and ensuring information interaction through the centralization system with the upper level control and management systems.

2. The integrated control system of the rolling stock technical condition on the move of the DISK-2 train subsystems consists for detecting overheated axle boxes, braked wheels, dragging parts, wheel irregularities in the course of rolling, deviations of the rolling stock's upper dimension, overload or uneven loading of the load.

3. The automated contactless complex for the rolling stock wheelsets control is intended for contactless control and analysis of rolling stock wheelsets parameters.

4. The automated control system of the SAKMA automatic coupling mechanism is used to monitor the faults presence in automatic coupling devices, due to which automatic coupling of freight cars self-uncoupling may occur on the train course.

5. The detector of defective DDC wheels belongs to the floor means of cars technical condition automatic diagnostics while the train is in motion and is designed to detect wheelsets with defects on the wheels rolling surface causing unacceptable cars and the track unsprung elements dynamic overload.

6. Automated diagnostic system for measuring cars wheel pairs on the approaches to the station is designed to measure the rolling surface geometrical parameters, as well as identify wear and defects of solid wheels on the train, detect wheelset faults and promptly transfer the received information to the nearest VET.

II. ANALOG SIGNAL PROCESSING

In the modern world, with significant development of computing technology, it has become much more productive to process the signal with its help. Before using a computer to process a signal, it must be digitized. Usually, the original signal received is in analog form. The process of signal transition from analog to digital is called analog-to-digital conversion, and the device for this procedure is called analog-to-digital converter (ADC). During digital signal processing using the principle of decomposition into components. It's possible to realize the decomposition of a signal into a linear combination of signals as well as into complex shapes, such as a set of sinusoids with a discrete Fourier transform.

An alternative process is called signal convolution. Two components can be distinguished in it: noise (interference) and informationally significant. Interference may appear in the signal under different conditions. Quantization noise is very common. This type of interference is caused by an error while measuring the continuous signal level with its discrete value.

Another reason for the appearance of noise may be a distortion of the hardware (sensors) with which ones the signal was established. Interference is negative because it distorts the necessary information, reducing the quality of diagnosis. If we consider the signal of the instrument analysis of crane metal structures, it is noticeable that in the absence of a violation, the signal from the devices has a small amplitude. This means the presence of noise. If the noise value is far from the average values, then this will distort the measurement picture, therefore, the interference variance is it's characteristic.

If we consider in detail the signal from the instrument analysis of crane metal structures in the presence of damage, then the presence of defects can be determined by sharp jumps in the signal. It's clear that the maximum extremum of the signal amplitude characterizes its informationally significant part. The signal-to-noise ratio is an abstraction of a quantity that characterizes the signal quality (no noise). As this value increases, the intensity of the signal at the noise level increases proportionally and versa vice. In most cases, this value is measured in Bel (in a logarithmic scale). The signal-to-noise ratio is determined by the signal-to-noise ratio. This parameter is an indicator of the quality of the signal from the device, examining the state of the metalwork, and characterizes its quality factor. In total, with criterias that are also relevant to the study, this measure may be the basis for choosing the best method for diagnosing metal structures [2].

There is also another statistical model of the system for preventing the metal structures critical state. It is

implemented in the information support system for optimizing the periodicity of monitoring the lifting machines metal structures state. The system involves four options for work: the calculation of prevention system of metal structures indicators; optimization of the preventive metal structures recovery period; critical level optimization of the control platform linear size and joint parameters optimization of the prevention regime. While making the initial parameters necessary for the system to work, the user will be able to note the need for registration while developing a preventive recovery mode for calendar time. The result of the system is the ability to see on the computer screen and, by necessary, save to the user-specified file [3].

III. DEFECT DETECTION SYSTEMS

Investigating the operation of metal structures with a variable load over time, it should be noted that its critical state is the exhaustion of crack resistance and metal fatigue. Damage due to metal fatigue manifests itself in the surface and near-surface layers of metal loading. At the first stages, it's possible to detect them and get a quantitative assessment with the help of special equipment. In the following steps (the French line) this can be realized only with the help of magnifying devices. Each stage of fatigue corresponds to different changes in the smoothness of the surface. The slip lines manifestation is the first stage. The formation of critical size macrocracks is the final one. On this basis, it can be noted that a quantitative and changes qualitative assessment in the smoothness of the surface can provide the necessary data on the fatigue damage degree of the product under study, and, of course, determine its life before failure. It's possible to determine the degree of changes in the surface relief by direct measurement, as well as by changing the optical properties. This is produced by using scanning tools by means of reflectometric methods for diagnosing the metal structure state [4]. The use of the reflectometric diagnostic method makes it possible to reduce the influence of the human factor on the diagnostic procedure, and, as follows, to increase the objectivity of the assessment. In addition, the use of the reflectometric method makes it possible to automate the monitoring process with the least cost. This allows you to reduce the time spent on diagnosis and increase the objectivity of the metal structures residual life assessment.

Using the shape parameters of the surface scattering indicatrix is one of the implemented options in the fatigue damage detection system with linear correction. The principle of the system is to control the conversion ratio of the processing channel in the area components that is a control channel as a scattered surface. As a function of the other component current values through the processing channel that is a correction channel. The emitter 1 and the slit diaphragm 2 direct the light beam to the surface of the control platform 8 and form a light mark on it (Fig. 1). In the process of scanning by photodetectors 3 and 4, the specular and diffuse components of the light scattered by the surface are analyzed. The mirror component channel is the control one, and the diffuse component channel is corrective. The diffuse channel component performs linear correction of the control channel transformation function, depending on the brightness in this channel.

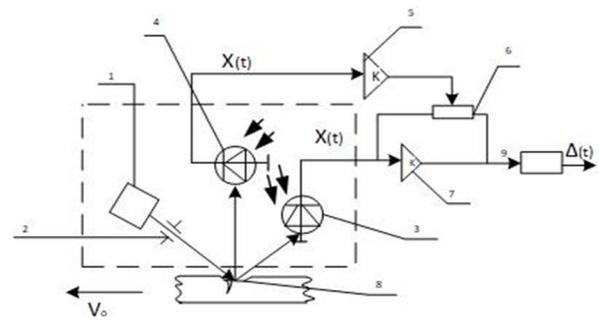


Fig. 1. Fatigue damage detection system with linear mirror gain correction

If surface defects of the mirror channel luminance were not detected, they are caused by noise from colour surface areas inhomogeneity, extraneous illumination, fluctuations in the degree of medium compensated transparency by adjusting the gain of the control channel amplifier 7 by changing the resistance value of the resistor 6 diffuse brightness fluctuations in the correction channel by the amplifier 5 output signal. If there is a mechanical defect, then due to energy redistribution within the scattering indicatrix, specular brightness component decreases and increases diffuse. At the same time, the gain of the control channel is also reduced. A decrease in the control channel gain caused by an increase in the diffuse component causes a further reduction in the signal, which is turn increases the control resolution while reducing the noise influence [5].

The defect detection system (Fig. 2) works in the same way, in which the diffuse component channel of the scattered light is the control channel, and the mirror component channel is the correction channel.

The manifestation of a defect, in this case, causes an increase in the control channel amplifier gain, as well as an increase in its output signal, which is more convenient for recording than when the signal is reduced.

Due to the fact that the correction channel gain is a light flux diffuse component luminance function scattered by the surface, the value of the information signal will depend on the presence of mechanical defects on the monitored surface, as well as its microgeometry.

In system on fig. 3, an additional resistor 10 is introduced in the correction channel feedback circuit amplifier 7, the resistance of which changes under the same amplifier output signal control [6].

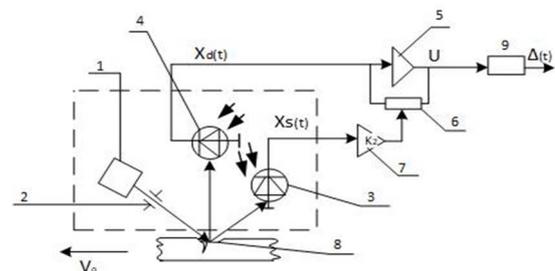


Fig. 2. Fatigue damage detection system with linear correction of diffuse channel amplification

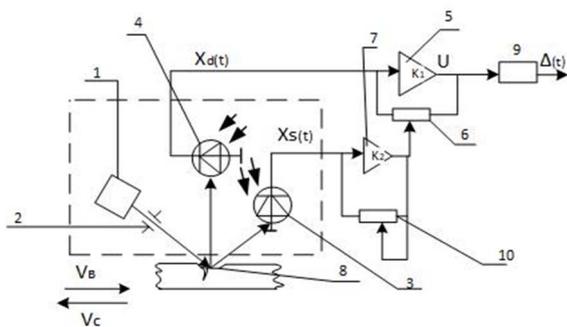


Fig. 3. Invariant fatigue damage detection system with control channel gain correction

In the areas of maximum possible destruction in the zone of the alleged development and passage of the crack, control platforms are being prepared. The reference site is an area of the surface that has been treated with high purity. The cleanliness of the control platform surface depends on the resolution of the means used for scanning. Scanning the test platform surface is carried out by optical reflectometric sensors that implement the principles of fixing the parameters of optical radiation scattered by the surface being monitored. It also allows you to show the existing changes in the optical properties of the surface.

During the test platform scanning process, the sensors evenly move in a direction that is perpendicular to the expected growth of the macrocrack, so that in the process to cross the alleged crack. After the scanning process, a reverse movement is performed to the sensor starting point.

The test platform volume is chosen so that in the study course, the optical sensor light mark ensures that the dimensions of the estimated crack maximum development occur.

The process of loading cycles in the metal structure under investigation is carried out with the accumulation of fatigue damage.

Such damage manifests itself in reflection on the surface or near-surface layers of metal while changing its topography and structure. The change in the microrelief and relief, firstly, influences the change in the optical properties of the surface, which in turn is recorded using an optical reflectometric sensor [7].

The automated system for monitoring the state of the metal structure, which has been developed, two main parts consists: a drive and an optical sensor.

A reflectometer sensor 3d model is presented, which differs from the existing ones by smaller dimensions and high measurement accuracy. This was achieved due to the developed optical system and printed circuit boards, the use of the latest optoelectronic and semiconductor elements. The development was carried out in Autodesk Inventor CAD — an Autodesk three-dimensional solid-state and surface design system designed to create digital prototypes of industrial products [8].

The radiation source is a surface-emitting laser with a vertical resonator (VCSEL) with a microlens integrated into the housing. Compared with traditional lasers, the main positive properties of a VCSEL laser include low angular divergence and an output optical radiation symmetrical

radiation pattern, temperature and radiation stability, group manufacturing technology and the ability to test instruments directly on the plate. A microcontroller that provides signal removal from photodetectors controls a stepper motor. On the printed circuit board where it is located, digitized and transferred to the industrial computer for the following analysis. Transmission is carried out through the RS-232 interface.

To scan the metal surface, the sensor moves along the monitored surface. The sensor and the drive control, as well as the obtained data transfer to the PC, is carried out using a microcontroller installed in the sensor [9].

If the equivalent state variable of operation isn't constant, its functioning is expressed in non-constant noise at the input of the reflectometric sensor. They are caused by the presence on some areas of surface study with several different classes of surface finish. The boundaries of each of these surfaces are surrounded by irregularities having a single scale. Monitoring the flow of parts with different irregularity scales poses the choosing appropriate control algorithm problem for conditions that are determined by the current value.

Such a task could be solved by a method of detecting defects on the product surface, which implements the corrective control strategy idea (Fig. 4).

On the surface of the monitored product 8, which is rotated, a light beam is emitted from the emitter 1, which is formed by the diaphragm 9. The photodetectors 2 and 3 fix the specularly and diffusely reflected light fluxes at points at equal distances. It excludes the existence of an influence on their amplitudes ratio losses in the field of scattering indicatrix form propagation and accurate approximation, from the point of light flux incidence on the product and between themselves.

The reflected light fluxes amplitudes at each registration point, converted into electrical signals by photodetectors, are equal in time and compared with the amplitudes of the signals from the light fluxes at the other points in the converter - comparator 4. According to the results of the amplitude comparison, the computing unit 5 selects a program from block 6 of permanent memory, which determine the defects on the product surface [10,12].

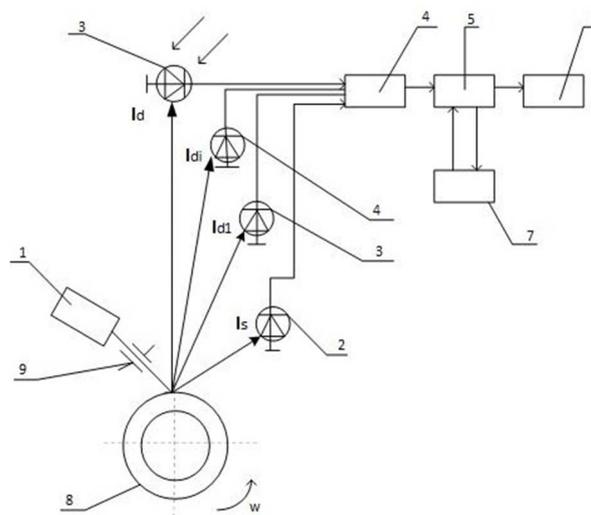


Fig. 4. The method of detection of surface defects with switching algorithms

The ability to most accurately determine the position of a defect and the degree of its development gives us a huge advantage over the rate of change in the state of the metal structure. Simple working equipment will decrease, the results of work will increase, it will be possible to send equipment less frequently for urgent repairs.

REFERENCES

- [1] A. Lay-Ekuakille, N.I. Giannoccaro, L. Spedicato, P. Vergallo, A. Massaro, R. Cingolani, A. Athanassiou New generation of optical robotic sensor applied to small notch detection // 2011 Fifth International Conference on Sensing Technology, Palmerston North, New Zealand, 28 Nov.-1 Dec 2011, DOI: 10.1109/ICSensT.2011.6136982.
- [2] Y. Gotoh, H. Hirano, M. Nakano, K. Fujiwara, N. Takahashi Electromagnetic nondestructive testing of rust region in steel // IEEE Transactions on Magnetics Volume 41, Issue 10, Oct. 2005, pp. 3616 – 3618.
- [3] Václav Triska, Ondřej Flášar Non-destructive inspection of composite specimen with integrated lightning protection using pulsed thermography // 2017 International Conference on Military Technologies (ICMT), Brno, Czech Republic, 31 May-2 June 2017, DOI: 10.1109/MILTECHS.2017.7988817
- [4] Shotaro Kawataki, Takayuki Tanaka, Satoru Doi, Shigeru Uchida, Maria Q. Feng Nondestructive inspection of voids in concrete by multi-layered scanning method with electromagnetic waves // 2017 IEEE International Conference on Mechatronics (ICM), Churchill, VIC, Australia, 13-15 Feb. 2017, DOI: 10.1109/ICMECH.2017.7921127.
- [5] Hui Min Kim, Gwan Soo Park A New Sensitive Excitation Technique in Nondestructive Inspection for Underground Pipelines by Using Differential Coils // IEEE Transactions on Magnetics, Volume 53, Issue 11, Nov. 2017, Article Sequence Number 6202604.
- [6] Yuji Gotoh, Makoto Tohara, Ryo Nakamura Electromagnetic Inspection for Detecting Defect of Underground Part of Road Sign Pillar // IEEE Transactions on Magnetics, Volume 54, Issue 11, Nov. 2018, Article Sequence Number 6202304.
- [7] Kai Zheng, Jie Li, Chun Lei Tu, Xing Song Wang Two opposite sides synchronous tracking X-ray based robotic system for welding inspection // 2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP) Nanjing, China, 28-30 Nov. 2016, DOI:10.1109/M2VIP.2016.7827334.
- [8] Shu-juan Wang, Xiao-yang Chen, Tao Jiang, Lei Kang Electromagnetic ultrasonic guided waves inspection of rail base // 2014 IEEE Far East Forum on Nondestructive Evaluation/Testing Chengdu, China, 20-23 June 2014, DOI:10.1109/FENDT.2014.6928248.
- [9] Nicola Ivan Giannoccaro, Luigi Spedicato, Aimè Lay-Ekuakille, Alessandro Massaro Automatic diagnostic by using a new optical sensor // 2015 IEEE Metrology for Aerospace (MetroAeroSpace) Benevento, Italy, 4-5 June 2015, DOI:10.1109/MetroAeroSpace.2015.7180665.
- [10] Gattiker A., Nigh P., Aitken R. An overview of integrated circuit testing methods. Microelectronics Failure Analysis. Desk Reference, 6 ed. ASM International 2011, Materials Park, Ohio 44703-0002, 9 p.
- [11] L. Su, T. Shi, Z. Liu, H. Zhou, L. Du, G. Liao Nondestructive diagnosis of flip chips based on vibration analysis using PCA-RBF. Mechanical Systems and Signal Processing, vol. 85, 2017, pp. 849–856.
- [12] E Zwicker, W Zesch, Roland Moser A modular inspection robot platform for power plant applications // 2010 1st International Conference on Applied Robotics for the Power Industry Montreal, QC, Canada 5-7 Oct. 2010 DOI: 10.1109/CARPI.2010.5624429.

Parametric Optimization Subsystem in LTspice Environment of Analog Microcircuits for Operation at Low Temperatures

Maxim V. Liashov
Department "Information Systems and
Radioengineering"
Don State Technical University
Rostov-on-Don, Russia
max185@mail.ru

Nikolay N. Prokopenko
Department "Information Systems and
Radioengineering"
Don State Technical University
Rostov-on-Don, Russia
prokopenko@sssu.ru

Andrei A. Ignashin
Department "Information Systems and
Radioengineering"
Don State Technical University
Rostov-on-Don, Russia
igan_96@mail.ru

Oleg V. Dvornikov
Minsk Research Instrument-Making Institute JSC (MNIPI JSC),
Minsk, Belarus
oleg_dvornikov@tut.by

Alexey A. Zhuk
Department "Information Systems and Radioengineering"
Don State Technical University
Rostov-on-Don, Russia
alexey.zhuk96@mail.ru

Abstract—A parametric optimization subsystem (PPS) has been developed that focuses on the use of the LTspice environment for designing low-temperature and radiation-hardened analog microcircuits in space instrumentation tasks. However, local optimization (NM, MAGPM), global optimization (DE, jDE, PSO, SA, ABC), multi-criteria optimization (NSGA-II, SPEA2, MO-CMA-ES), parallel optimization on a multi-core processor, distributed optimization on the cluster of personal computers, plotting the fitness functions and Pareto efficiency, operation with the latest version of LTspice are provided. An example of parameter optimization of the elements of a low-temperature output stage (BA) of an operational amplifier, realized on complementary junction field-effect transistors, is given. A comparison of the characteristics of optimal and non-optimal BA is presented.

Keywords—low-temperature electronics, junction field-effect transistors, optimization of analog electronic circuit, LTspice environment, buffer amplifier, operational amplifier

I. INTRODUCTION

The use of complementary junction field-effect transistors (CJFet) [1-4] is promising for constructing low-noise low-temperature and radiation-hardened analog microcircuits. This class IC circuitry is at the initial stage of development, because, due to the low mass production, many microelectronic firms did not pay sufficient attention to this sector of the electronic component base. At the same time, the extremely low noise level, as well as the CJFet technologies [5-7] developing in recent years, create initial conditions for the development of CJFet microcircuits operating in severe conditions (low temperatures, neutron fluxes, gammas, accumulated radiation dose, etc.).

Significant improvement of generalized (or priority) CJFet IC quality indicators can be provided through the use of special CAD systems that allow optimization of schemes in static and dynamic modes [8-10].

The purpose and novelty of this article is to describe a specialized parametric optimization system that focuses on operation in LTspice [11-14] using computer models of CJFet transistors [15], as well as to create the optimal CJFet

scheme of a buffer amplifier [7] for operation at low temperatures.

II. DESCRIPTION OF OPTIMIZATION SUBSYSTEM IN LTSPICE

To optimize analog circuits, incl. CJFet, this work uses a library for distributed evolutionary computations with open source code DEAP (Distributed Evolutionary Algorithm in Python) [16].

DEAP supports a number of bioinspired algorithms, such as: genetic algorithms, swarm algorithms, multi-criteria evolutionary algorithms NSGA-II, SPEA2, MO-CMA-ES and others. The library already contains most of the basic functions required for evolutionary computing, so its users can easily create different types of both single and multi-criteria evolutionary algorithms and execute them in parallel on multi-core processors and clusters. DEAP is ideal for rapid prototyping and can be used with a variety of other Python libraries for data processing, as well as other machine learning methods.

The structure of the developed optimization subsystem for the LTspice environment is shown in Fig. 1.

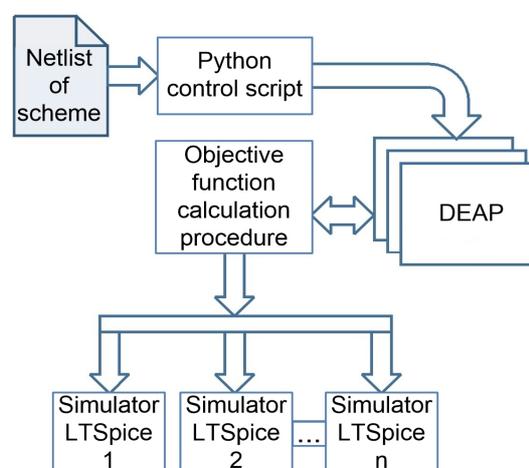


Fig. 1. The structure of the parametric optimization subsystem

The study has been carried out at the expense of the grant from the Russian Science Foundation (Project No. 16-19-00122-P).

A scheme description file in the form of a net-list and a script tasks come on the subsystem's input for its optimization in Python. The script calls the optimization method from the DEAP library.

To calculate the objective function of the optimization algorithm, the configuration of the scheme and the commands for calculating the indicators are formed, and the LTSpice circuit simulator is started. In this case, several processes can be specified in the script for optimization on multi-core personal computers (PC), which will significantly speed up the work of the algorithm. After the simulation is completed, the calculated indicators are used to determine the objective function. At the end of the optimization process, the subsystem displays the optimal values of the circuit parameters and the obtained indicators, and also generates the optimization process graphs.

The developed subsystem of analog and analog-digital circuit optimization for the LTSpice environment solves the following tasks: Local optimization (NM, MAGPM); Global optimization (DE, jDE, PSO, SA, ABC); Multi-criteria optimization (NSGA-II, SPEA2, MO-CMA-ES); Parallel optimization on a multi-core processor; Distributed optimization on a PC cluster; Plotting of fitness function and Pareto efficiency; Work with the latest version of LTSpice-17.

III. EXAMPLE OF PARAMETER OPTIMIZATION OF THE BUFFER AMPLIFIER ELEMENTS ON COMPLEMENTARY FIELD-EFFECT TRANSISTORS IN LTSPICE

Fig. 2 shows the original scheme of the buffer amplifier (BA) [7] on complementary field-effect transistors (CJFET, OAO Integral, Minsk), which was used to demonstrate the capabilities of the developed optimization subsystem.

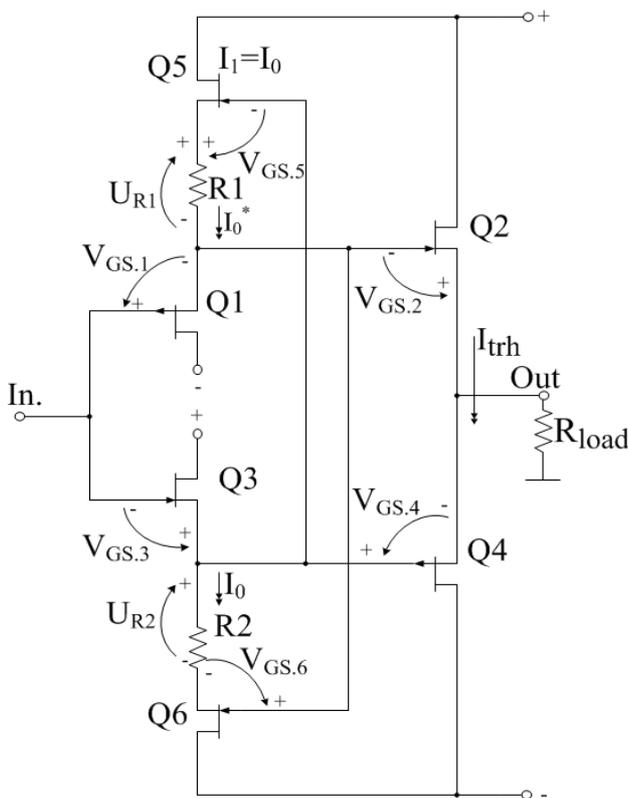


Fig. 2. CJFet buffer amplifier for operation at low temperatures

The static mode of the BA circuit elements is determined by resistors R1, R2, as well as the parameters of the drain-gate characteristics of field-effect transistors Q1-Q4, the length and width of their channel.

The input data of the optimization subsystem include a scheme description file in the form of a net-list and a control script for optimization in the Python language. The scheme description file corresponds to the LTSpice circuit simulator used for calculating parameters [11-14].

Fig. 3 shows a fragment of the description file of the BA circuit, which includes objective variables.

```

*** measurements
.meas DC tVbias FIND V(out)-V(in1) AT 0
.meas DC tVmax FIND V(out) AT 5
.meas DC tVmin FIND V(out) AT -5
.meas DC tIcon FIND I(V1) AT 0
.meas DC Vabs PARAM tVmax-tVmin
.meas DC Icon PARAM tIcon
.meas DC Vbias PARAM tVbias

```

Fig. 3. Fragment of the description file

In this example, the following notation is introduced:

- Vbias - offset voltage of the BA, which, according to the conditions of the specific task, should be minimized (brought nearer to zero value);
- Icon - total static current consumption of the BA, which, when the first condition is met, should be as low as possible;
- Vabs - range of change in the output voltage of the BA, which at the specified supply voltage ($\pm 5V$) should have the highest value (close to 10 V).

Fig. 4 presents a fragment of the script in the Python language in which the following is indicated:

- parameters of the algorithm for optimization;
- optimizable BA scheme;
- calculated metrics;
- restrictions for optimizable parameters.

In this example, the parameters of the following elements are optimized:

- R_1, R_2 – resistance of resistors R1, R2. It can be changed within the range of $1k \div 300k$. This changes the current consumed by the circuit BA in static mode and dynamic parameters. In particular cases, R1 and R2 may not be the same.
- N_2, N_4 – the number of parallel-connected field-effect transistors Q2, Q4. It can be changed within the range of integer numbers from 1 to 10. It is also possible to change the length and width of the channels Q2, Q4.
- R_load – load resistance of the BA. In this case, this parameter can take the following values: 1k, 10k, 100k.

```

# -----
#           Algorithm Parameters
# -----

# Number of parallel processes for
# optimization
processes = 5

# Parameters for evolutionary algorithm
weights = (1.0, 1.0, -1.0)
pop_size = 50
max_gen = 300
cx_prob = 0.8
mut_prob = 0.2

# Optimizable file. Net-list
fname = 'schem356.cir'

# Calculated metrics
metrics = ['vabs', 'icon', 'vbias']

# Restrictions for optimizable parameters
param = {
    'R_1': [1000, 300000],
    'R_2': [1000, 300000],
    'N_2': [1, 10],
    'N_4': [1, 10],
    'R_1': [1000, 100000]
}

```

Fig. 4. Fragment of the script in the Python language to set the parameters of the algorithm for optimization

IV. RESULTS OF SIMULATION EXPERIMENT

Before the optimization the BA scheme of Fig. 2 was characterized by the following parameters:

- offset voltage - 120 mV;
- static current consumption - 200 μ A;
- maximum change in the output voltage - 8.1 V.

The offset voltage of the BA (excluding manufacturing tolerance of element parameters) Vbias was selected as a priority parameter that should be optimized, which determines the static parameters of operational amplifiers with current negative feedback [17,18].

Since there is more than one optimizable parameter, the selection of elements of the BA scheme on CJFets is a multi-objective optimization task. To solve this problem, the multicriteria optimization NSGA-II algorithm from the DEAP library [19] was applied. Computer models of JFet transistors, considering their behavior at cryogenic temperatures were used [15].

The optimization process graphs for three parameters of the BA scheme are shown in Fig. 5.

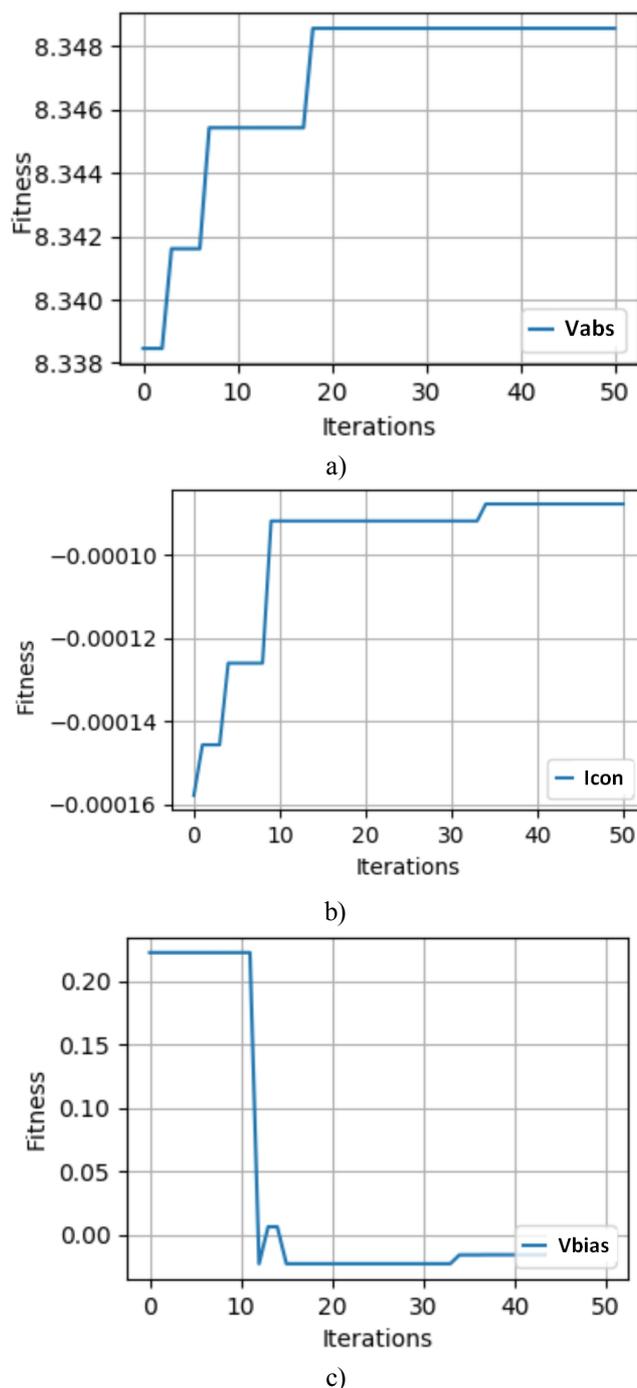


Fig. 5. Optimization process graphs for Vabs (a), Icon (b) и Vbias (c)

From the presented graphs it can be seen that the optimal parameters were found in 50 iterations of the evolutionary algorithm. The following optimal values were obtained:

- R_1 = 8847 Ohm;
- R_2 = 7968 Ohm;
- N_4 = 9;
- N_2 = 10;
- R_1 = 100 kOhms.

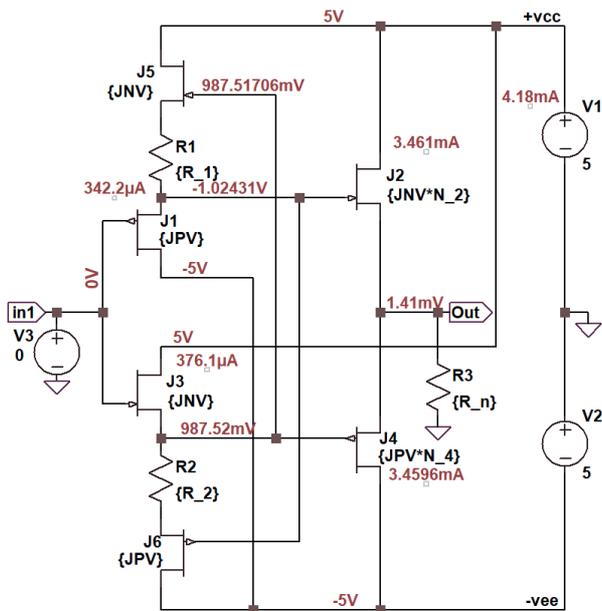


Fig. 6. Static mode of the optimal BA scheme of Fig. 2 at a temperature of 27 °C

Fig.7 shows the dependence of the output voltage of the BA from Fig. 2 on the input voltage for the optimal parameters found.

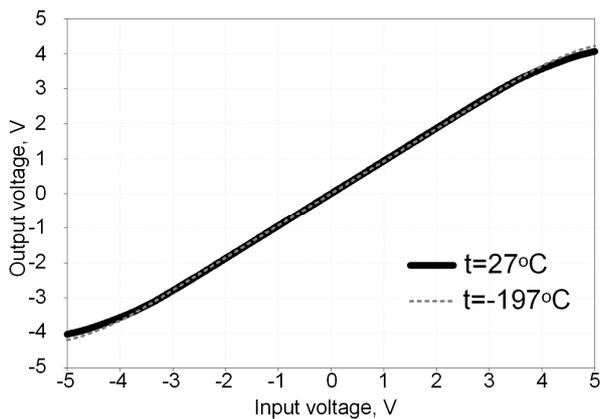


Fig. 7. The dependence of the output voltage of the BA on the input voltage at a temperature of $t = 27^{\circ}\text{C}$ and $t = -197^{\circ}\text{C}$

In this case, the numerical values of the BA parameter, optimized by the magnitude of the offset voltage V_{bias} are:

- offset voltage $V_{\text{bias}} = 1.49 \text{ mV}$
- range of change of Ba output voltage $V_{\text{abs}} = 8.38 \text{ V}$
- Static current consumption $I_{\text{con}} = 4.17 \text{ mA}$

Thus, the BA scheme optimization of Fig. 2 enabled to improve the offset voltage more than 100 times.

V. CONCLUSION

The advantages of the developed parametric optimization subsystem, which is aimed at using the LTspice environment for designing low-temperature and radiation-resistant analog microcircuits, are demonstrated by the example of the optimal choice of parameters of the low-temperature CJFet output stage of the operational amplifier. This advantage has provided a reduction in the systematic component of the zero bias voltage BA more than 100 times.

REFERENCES

- [1] J. Caldwell, "Distortion and source impedance in JFET-input op amps," Analog Application Journal, 4Q, 2014, pp. 4-6
- [2] M. Snoeij, "A 36V 48MHz JFET-Input Bipolar Operational Amplifier with 150 μV Maximum Offset and Overload Supply Current Control," ESSCIRC 2018 – IEEE 44th European Solid State Circuits Conference, pp. 298-301. DOI: 10.1109/ESSCIRC.2018.8494262
- [3] JoAnn P. Close, F. Santos, "A JFET input single supply operational amplifier with rail-to-rail output," 1993 Proceedings of IEEE Bipolar/BiCMOS Circuits and Technology Meeting, 4-5 Oct. 1993, Minneapolis, pp. 149-152. DOI: 10.1109/BIPOL.1993.617487
- [4] M. Snoeij, M.V. Ivanov, "A 36V JFET-input bipolar operational amplifier with 1 $\mu\text{V}/^{\circ}\text{C}$ maximum offset drift and -126dB total harmonic distortion," 2011 IEEE International Solid-State Circuits Conference, 20-24 Feb. 2011, San Francisco, CA, USA, pp. 248-250. DOI: 10.1109/ISSCC.2011.5746305
- [5] D. G. Drozdov, "Microwave complementary bipolar technological process with a high degree of symmetry of the dynamic parameters of transistors," Abstract of Ph.D. dissertation, Moscow, 2017, 17 p. (in Russian).
- [6] O. V. Dvornikov, *et al.*, "Cryogenic Operational Amplifier on Complementary JFETs," 2018 IEEE EWDTs, Kazan, 2018, pp. 1-5. DOI: 10.1109/EWDTs.2018.8524640
- [7] N. N. Prokopenko, *et al.*, "Buffer amplifier based on complementary field-effect transistors with p-n junction control for operation at low temperatures," Patent appl. RU 2019118999, June 19, 2019 (in Russian)
- [8] M. M. Gourary, S. G. Rusakov, S. L. Ulyanov and M. M. Zharov, "Optimization approach to design of linear voltage regulators for system on chip," 2017 SIBCON, Astana, 2017, pp. 1-4. DOI: 10.1109/SIBCON.2017.7998578
- [9] D. M. Binkley, *et al.*, "Optimizing Drain Current, Inversion Level, and Channel Length in Analog CMOS Design," Analog Integrated Circuits and Signal Processing, 47(2), pp. 137-163. DOI:10.1007/s10470-006-2949-y
- [10] M. d. Hershenson, S. P. Boyd and T. H. Lee, "Optimal design of a CMOS op-amp via geometric programming," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 20, no. 1, pp. 1-21, Jan. 2001. DOI: 10.1109/43.905671
- [11] LTspice® XVII, 1998-2019 Analog Devices Corporation All rights reserved [Online]. Available: <http://LTspice.linear.com> Accessed on: May 13, 2019
- [12] LTspice Tutorials, http://www.simonbramble.co.uk/lt_spice/ltspice_lt_spice.html Accessed on: May 13, 2019
- [13] More LTspice, Lab 2 [Online]. Available: http://faculty.engineering.asu.edu/eee202lab/wp-content/uploads/2017/01/2_More_LTspice.pdf Accessed on: May 13, 2019
- [14] An Introduction to LTSpice [Online]. Available: <https://forum.digikey.com/t/an-introduction-to-ltspice/2023> Accessed on: May 13, 2019
- [15] O. V. Dvornikov, *et al.*, "The accounting of the simultaneous exposure of the low temperatures and the penetrating radiation at the circuit simulation of the BIJFET analog interfaces of the sensors," 2017 SIBCON, Astana, Kazakhstan, 2017, pp. 1-6. DOI: 10.1109/SIBCON.2017.7998507
- [16] De Rainville, *et al.*, "DEAP: A Python framework for Evolutionary Algorithms," GECCO'12, pp. 85-92. 10.1145/2330784.2330799.
- [17] M. Djebbi, A. Assi and M. Sawan, "An offset-compensated wide-bandwidth CMOS current-feedback operational amplifier," CCECE 2003 (Cat. No.03CH37436), 2003, pp. 73-76 vol.1. DOI: 10.1109/CCECE.2003.1226347
- [18] N.N. Prokopenko, *et al.*, "Maximum rating of Voltage Feedback and Current Feedback Operational Amplifiers in Linear and Nonlinear Modes," Proceeding of the ICCSC'06, Politehnica University, Bucharest, Romania: July 6-7, 2006, pp.149-154
- [19] K. Deb, *et al.*, "A fast and elitist multi-objective genetic algorithm," NSGA-II. IEEE Transaction on Evolutionary Computation, 6(2), pp. 181-197.

Boosting Model of Bioinspired Algorithms for Solving the Classification and Clustering Problems

Ilona Kursitys
Computer Aided Design
Southern Federal University
Taganrog, Russia
i.kursitys@mail.ru

Alexander Natskevich
Computer Aided Design
Southern Federal University
Taganrog, Russia
natskevich.a.n@gmail.com

Elvira Tsyruelnikova
Computer Aided Design
Southern Federal University
Taganrog, Russia
ehbolshova@yandex.ru

Abstract— The paper considers the methods of boosting models application for solving clustering and classification problems. The main characteristics of boosting models are analyzed in the paper. The authors present the classification and clustering problems statement, describe popular modern and classical algorithms and analyze their benefits and shortcomings. According to the conducted research, a modified boosting model to solve the classification and clustering problems is developed and presented in the paper. The authors also compare the approaches of boosting and bagging and demonstrate their strengths and weaknesses. The paper describes the algorithms to be used in the developed boosting model. A new model of solving optimization problems is based on the usage of a weighted set of bioinspired clustering algorithms and their boosting. The heuristic of the suggested boosting method involves the use of a probability matrix providing a weighted estimation of the results obtained by different learning algorithms to achieve the highest quality of the problem solution. The developed approach is based on the usage of weighted data sets containing the probability of adding each individual element in a particular cluster. The conducted experimental research has shown that the developed boosting approach allows us to obtain the solutions equal or superior to those obtained by the popular algorithms.

Keywords—boosting, clustering, classification, evolutionary modeling, swarm algorithms, machine learning, bioinspired algorithms.

I. INTRODUCTION

In terms of solving many scientific, social and business tasks, there is a need for solving data mining tasks, which becomes more difficult due to the fact that one of the most expressed trends of the society development is the constant increase of the semistructured data scope [1]. The IMB statistics from [2] shows that at least 2.5 exabytes of the data are generated a year [2].

The mentioned problem justifies the relevance of creating new scalable algorithms for data mining which can provide good clustering results in a reasonable time. One of the most used data mining methods is clustering, which is explained by the need for dividing the large growing data scope into the clusters [1] for the further simplification of its processing for retrieving information and solving different scientific tasks. Initially, there is a set of objects to be divided into a set of clusters in such a way that each individual group includes the most similar objects according to

the used measure. The number of clusters can be preset or defined in the process of the algorithm work. Each object can be included in each cluster.

Clustering can be considered as the most important and promising methods for unsupervised learning [3]. The clustering problem is referred to the NP-complete tasks, thus, the full enumeration methods are impossible to use, and the exact methods are weakly effective. Therefore, the development of effective methods for solving the clustering problem is relevant today.

To solve the mentioned task there are a lot of algorithms which differ from each other in time consumptions, algorithmic complexity, and different working conditions.

A lot of clustering methods were classified by such scientists as Donkuan, X. Yingjie T. whose research results are analyzed and shown in [4]. They divided the available clustering algorithms into two large groups: classical and modern methods.

Ensembles algorithms can be noted as one of the most effective among the modern methods. Ensembles algorithms include such approaches as algorithms based on the genetic approach [5-11] and algorithms based on the usage of the fuzzy sets. The main idea of such algorithms includes generating the set of initial results of clustering according to a certain method. The conclusive results of clustering are obtained by integration of the initial results of clustering solved by the different algorithms. The benefits of such a method include the opportunity to parallelize the used algorithms. The shortcomings involve the complexity of consensus function development [4]. The algorithmic complexity highly depends on the types of algorithms included in the ensemble.

Among the popular methods, we can distinguish boosting which is explained by its rather active development and high quality of the solutions. Boosting is also based on the ensembles method. The main procedure involves the sequential building of composition of machine learning algorithms. Each successive algorithm tends to compensate for shortcomings of all the previous compositions of algorithms to increase the effectiveness of solving a certain optimization task [4, 12]. For instance, in terms of the clustering task, boosting allows us to consider the specific organization of each individual data set and to choose the most effective algorithm for its processing.

The rest of the paper is organized as follows. In the next section, the paper presents the main ideas of boosting and bagging and the well-known approaches of using them. The third section is devoted to the classifications and clustering tasks and the objective function in order to assess the quality of the solution. The fourth section describes the developed boosting algorithm for solving the clustering task, its structure and main steps for the successful work. The fifth section presents the results of the experiments, which are based on the comparison of the developed boosting algorithm and the well-known Approximate kernel k-means (AKMM) algorithm. The last section summaries and concludes the paper.

The next section of the paper presents the boosting concept and its main ideas.

II. BOOSTING MODEL FOR SOLVING THE CLASSIFICATION TASK

Combining the ideas of several different algorithms is a popular method used for solving supervised and unsupervised learning tasks. Among the mentioned methods we can distinguish two popular effective approaches: bagging (bootstrap aggregating) and boosting. The idea of bagging includes building several independent models or algorithms to implement a common solution obtained by voting or averaging methods [13]. For instance, such approach is used in Random Trees and Random Forest algorithms. The idea of boosting is completely opposite: different models or algorithms are used sequentially to solve the required problem. In terms of the boosting approach, each following model can estimate and consider the results and mistakes of the previous model to improve the final solution.

In the beginning, boosting was created for solving the classification task. The idea of boosting involves the fact that combining several weak classifiers can provide more effective results than a single classifier. It should be noted that a weak classifier can add the input elements in the proper classes with significantly less error than the random classification (0.5 in binary case) [14].

In general, the boosting model, as well as the bagging model, are based on the idea of building the ensemble of algorithms. The mentioned ensemble is presented as an algorithm for solving classification or clustering tasks, including several weak algorithms (classifiers) [14]. The weak algorithms are combined to obtain a single strong algorithm. The way of combining the algorithms depends on the specific problem to be solved. Freund and Shapire describe this idea in [15], assuming that it is much easier and effective to train the ensemble of the simple (weak) algorithms than the difficult (strong) one.

For instance, instead of training a single large neural network we can train several smaller neural networks and then combine their results of solving a required task. The model of such an approach is demonstrated in the Fig. 1, where $H_m: X \rightarrow \{-1, +1\}$ is a binary classifier m (for $m = 1, \dots, M$) and $x \in X$ is an input element to be classified. Such an idea includes combining the solutions obtained by each individual classifier $H_m(x)$. The final solution of the ensemble is represented as $H(x)$ and can be obtained by different methods such as weighted voting or a simple majority.

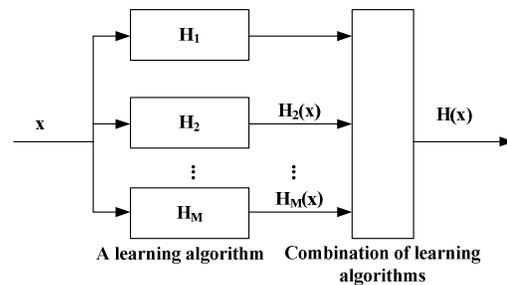


Fig. 1. General model of the ensemble of classifiers

The boosting model ideas involve multiple uses of weak algorithms to obtain the sequence of algorithms which results can be integrated into the final solution as demonstrated in Fig. 1. At each step of the algorithm, each obtained solution is estimated according to the accuracy of the solution given by each particular algorithm. This allows the boosting algorithm to focus on the data objects which were classified incorrectly. Some boosting algorithms can use particular criteria for selecting every weak algorithm, which can provide the solution of higher quality.

Freund and Schapire developed and described the adaptive boosting algorithm (AdaBoost) in [15, 16]. The main idea of AdaBoost includes using the weighted version of a certain set of elements, which is used multiple times. This assumes the large size of the mentioned set as it was in the algorithms described above. Nowadays, the AdaBoost algorithm is actively investigated and frequently used for building the ensembles of classifiers in a reasonable time. The algorithm trains a certain set of weak learning algorithms to build a weak classifier using the model demonstrated in Fig.1. The weak classifier is created by sequential application of the re-weighted data set containing the weights obtained in accordance with the accuracy of the results of the previous classifiers. Every time the algorithm uses the same data set with the entities weighted according to their correct or incorrect classification performed by the previously used classifiers. This allows the weak learning algorithms to focus on data which were classified incorrectly at the previous iteration. The problems of such an approach include the selection of the proper weak learning algorithm to obtain the base classifier in such a manner that the weight of the correctly classified objects is not reduced. If the base algorithm is accurate enough, it can start the classification process leaving the significant weight for the outliers and the noise entities to study them more precisely at the next iterations. The algorithm structure is demonstrated in Fig.2.

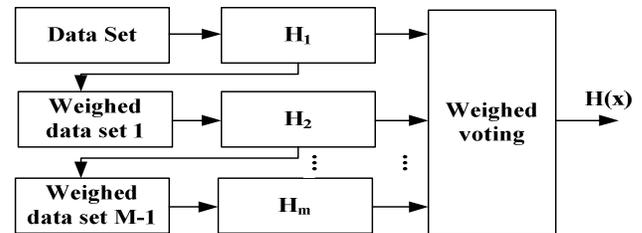


Fig. 2. Model of the AdaBoost algorithm

According to the analytical reviews provided by Dongkuan Xu, Yingjie Tian [4] and Ka-Chun Wong [1], a lot of standard boosting algorithms including AdaBoost, can give effective solutions of the classification task. However, we cannot adjust all modern available boosting algorithms to solving the unsupervised learning tasks. Thus, developing effective boosting algorithms for solving the classification and clustering tasks with polynomial complexity and appropriate time consumption is relevant today.

The next section presents the classification and clustering tasks.

III. ASSESSMENT OF THE BOOSTING MODEL

The boosting method is highly applied in the sphere of machine learning for solving supervised tasks such as the classification task. The core of such problems solution is to train a certain algorithm on the basis of previously classified data to create reliable predictions for unclassified data [14]. Supervised learning is a subdiscipline included in machine learning which also involves unsupervised learning based on the analysis of the unclassified data only.

On the basis of the approaches mentioned above the semi-supervised learning is distinguished. It includes the elements from both disciplines [18] since it is assumed that the data set involves the previously classified elements as well as the unclassified data. In [14] the authors also consider boosting algorithms focusing on solving these problems. One of the suggested solutions is adding the elements in the pseudo-classes.

This paper is based on the hypothesis that one of the possible solutions is reducing the classification task to the clustering task.

In terms of the supervised learning tasks, the algorithms are provided with a certain consensus function $h(\cdot)$, which consists of the solution of the classification task. The main purpose of solving the classification task is to categorize the objects into a predefined set of classes.

Generally, the outgoing data on the classification task solution is represented as Y , containing the information on two classes encoded as $\{-1, 1\}$. The main task of the machine is to learn on the basis of the training data set $(y_1, x_1), \dots, (y_n, x_n)$, which are classified for the further forecasting classification of new objects x_{new} . Forecasts of membership of the data x_1, \dots, x_n are represented as implementing from X , n is the size of the training data set. The machine task is to develop a forecasting rule $h(\cdot)$ for the correct classification of new data.

$$(y_1, x_1), \dots, (y_n, x_n) - (\text{supervised learning}) \rightarrow h(x_{new}) = y_{new} \quad (1)$$

In terms of the unsupervised learning, there are no training data sets and the consensus function is represented by the objective function of assessment of the obtained solution quality. The main task of the machine is to develop the initial rule of dividing the objects into clusters. The formula is represented in the following way:

$$(x_1, x_n) - (\text{unsupervised learning}) \rightarrow (y_1, x_1), \dots, (y_n, x_n) \quad (2)$$

In terms of solving the clustering task, we can assess the average inter-cluster distance or the average intra-cluster distance.

The solution of the clustering task is the set $V = \{Y^j | j=1, 2, \dots, k\}$. The planned variant of the solution V is the partition of the set of objects into a set of clusters.

Estimation of the solution V is represented by the objective function written as follows:

$$F = \frac{P^o}{P^i} \rightarrow \max, \quad (3)$$

Where P^o is the average inter-cluster distance, P^i – is the average intra-cluster distance.

Let us consider the mechanism of boosting organization with the use of a bioinspired algorithm.

The formula for calculating the intra-cluster distance is written as follows:

$$P^i = \frac{1}{X} \sum_{j=1}^n \sum_{i=1}^n p(x_i, c_j) \rightarrow \min, \quad (4)$$

where p is the distance calculated by the chosen metric, $x \in X$ is the current element, $c \in C$ is the centroid of the cluster, k – is the total number of elements, l is the number of elements in a certain j cluster.

The average inter-cluster distance describes the distance between the objects belonging to different clusters and is determined according to the following formula:

$$P^o = \frac{1}{U} \sum_{u \in U} p(u_i, u) \rightarrow \max, \quad (5)$$

where p – is the distance according to the chosen metric, u_i – is the concerned centroid, u – is the centroid regarding which the average inter-cluster distance is calculated, n – is the total number of clusters.

Thus, the boosting procedures can be effectively used for solving classification as well as clustering tasks. The authors in [12], demonstrates the high effectiveness of using the boosting algorithms for solving the semi-supervised tasks.

The next section describes the developed boosting algorithm for solving the clustering task.

IV. BOOSTING ALGORITHM FOR SOLVING THE CLUSTERING TASKS

In this paper, the authors use the model of boosting of bioinspired algorithms for solving the clustering task. The algorithm is based on the AdaBoost (Adaptive Boosting) algorithm mentioned above. The main idea of the developed algorithm is to use the weighted version of a certain set of algorithms and a set of probabilities determining the including

of each object into a concrete cluster. This set is used repeatedly allowing us to find the best algorithm for clustering each concrete data set. Fig. 3 demonstrates the algorithm work.

Let us consider the detailed work of the algorithm. At the first stage, the clustering algorithms are selected randomly. The solutions are estimated in accordance with the objective function.

Input parameters: Data set consisting of a set of objects to be clustered $X = \{x_i \mid i=1,2,\dots,n\}$, where n is the number of objects to be clustered. The set of the algorithms which can be used for clustering is $A = \{a_j \mid j=1,2,\dots,m\}$, where m is the number of the algorithm. T is the number of iterations.

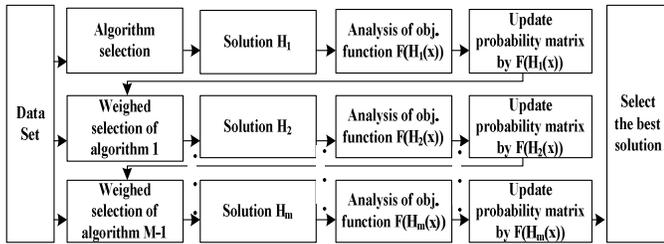


Fig. 3. Chart of the developed boosting model

One of the algorithm features is that the set of the algorithms used in the boosting process includes the bioinspired algorithms. Thus, it is possible to use the set determining the probability of including each element into a cluster which can be used in terms of ant colony algorithm.

The selection of each bioinspired algorithm is implemented by the following formula:

$$f_j = \left(\alpha \frac{V_b}{V_t} + \beta u_j \right), \quad (6)$$

where V_b is the value of the best solution objective function, V_t is the value of the current algorithm objective function at the current iteration, $u \in U$ is the probability of selecting the i -th algorithm from the set of the algorithms, which can be used for clustering, α is the coefficient determining the importance of the improvement criteria of the objective function, β is the coefficient determining the weight of the probability criteria for selecting each concrete algorithm.

In the process of obtaining the final solution by each bioinspired algorithm, the probability matrix of including each element from the data set in a cluster is updated for the sequential improvement of the solution by each following algorithm. The weights are updated according to the following formula:

$$f_{i,j} = (\alpha d(x_i, y_j) + \beta \tau_{i,j}), \quad (7)$$

Where α is the coefficient determining the weight of the criterion of distance between the object and the cluster centroid, d is the distance considering the used metric, $x \in X$ is the current object i , $y \in Y$ is the cluster j , β is the coefficient determining the weight of the criterion of probability of including the element

into the cluster j , $\tau_{i,j}$ is the occurrence threshold of including the element into a cluster.

The formula for calculating the probability $p_{i,j}$ of including the object x_i into the cluster $y_j \in Y$ is represented in the following way:

$$P_{i,j} = \frac{(\alpha d(x_i, y_j) + \beta \tau_{i,j})}{\sum_{j=1}^n (\alpha d(x_i, y_j) + \beta \tau_{i,j})}. \quad (8)$$

The algorithm consists of the following steps:

- 1) To initialize a set of probabilities of selecting each algorithm of clustering $U = \{u_j \mid j=1,2,\dots,m\}$.
- 2) While $t < T$
 - a) To run one of the clustering algorithms. To select the algorithm according to the formula (6)
 - b) To perform the clustering with the use of the selected algorithm, then to estimate the clustering quality according to formula (3), to estimate the intra-cluster and intercluster distance according to the formulas (4) and (5) respectively.
 - c) To calculate the weights of the probability of including each element into a cluster according to the results of the chosen algorithm work with the use of the formula (8)
 - d) To update the selection probability for the concrete selected algorithm.
 - e) To increment the number of iteration $t \leftarrow t+1$
- 3) To finish the cycle.
- 4) To select the best solution according to the objective function.

Step (a) includes the formula to estimate the quality of the concrete clustering algorithm. Since there is no data about the previous estimation at the first iteration, the formula is shown as follows:

$$f_j = (\beta u_j). \quad (9)$$

The next section demonstrates the results of the experimental research proving the effectiveness of the developed algorithm.

V. EXPERIMENTAL RESEARCH.

The purpose of the experiments is to determine the boosting model effectiveness for solving the clustering task. To check the algorithm, we propose to use the benchmarks with known optimal values. The first step is to investigate the controlling operators and their impact on the solution, such as the number of boosting iterations, values of the weights in the probability matrixes of choosing algorithms and including the elements into the clusters. To calculate the clear assessment of the model we conducted a set of experiments.

The time complexity of the boosting algorithm depends on the time complexity of the algorithms used in the boosting model. The initial time complexity of the boosting algorithm is $O(n)$. The experiment results demonstrate that in 98% iterations the solution space contains the optimal solution.

The parameters of boosting were set as follows: the number of boosting iterations is 50; the number of iterations for each individual algorithm affiliated with the ensemble is 100. We used the following algorithms in the boosting model: ant colony optimization (ACO) algorithm [18], artificial bee colony optimization (ABC) algorithm [19], genetic algorithm [20], and firefly colony optimization (FCO) algorithm [21].

To compare with the boosting model, we used several classic algorithms. Let us consider each of them.

Approximate kernel k-means (AKKM) is the clustering method based on the combination of the kernel method and k-means. The algorithm consists of two phases: kernel computation and clustering. The feature of the algorithm is the modified kernel matrix. In [22] the authors suggest partitioning of the matrix into several submatrixes to simplify the process of the kernel calculation. The time complexity of such an algorithm is based on the analysis of the kernel matrix building phase and the estimation of the clustering complexity. The formula for time complexity can be denoted as $O(m^3 + m^2n + mnCl)$, where m denotes the number of the elements for building the initial kernel matrix ($m < n$) and n denotes the number of the elements for clustering (m and n define the matrix dimension), C denote the number of clusters to partition the elements into, l denote the number of iterations. Table 1 demonstrates the results of the experiments in terms of time complexity.

TABLE I. TIME COMPLEXITY OF THE ALGORITHMS

The number of elements	Kernel calculation time (AKKM method)	Clustering time (AKKM method)	Clustering time (Boosting model)
100	1.40	17.70	17.30
200	1.64	22.57	21.64
500	3.82	28.56	26.48
1000	11.14	55.01	52.05
2000	22.80	134.68	131.86
5000	64.11	333.31	329.34

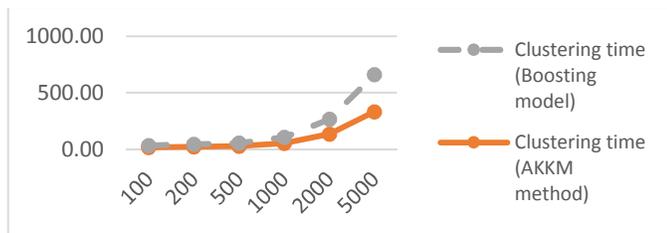


Fig. 4. Time complexity of the algorithms

The developed boosting model was compared with the popular algorithms such as k-means algorithm and AKKM algorithm according to the errors they can give with respect to the ICI (Incorrectly Clustered Instances). The experiments were carried out on the basis of the well-known benchmarks Ionosphere and Iris with different number of clusters. The percent of the objects which are clustered incorrectly in terms with the best solution is demonstrated in Table 2.

TABLE II. QUALITY OF THE SOLUTIONS OBTAINED WITH THE COMPARED ALGORITHMS

Dataset	The number of clusters	k-means	AKKM	Boosting model
Ionosphere	10	28.5	17.6	9.2
Ionosphere	20	26.3	16.5	6.4
Iris	10	22.3	9.3	5.7
Iris	20	18.6	7.3	4.4
Iris	30	15.1	5.3	2.1

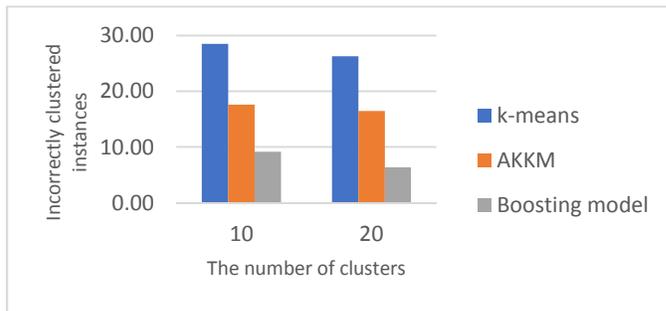


Fig. 5. Comparison of the quality of the algorithms (Ionosphere benchmark)

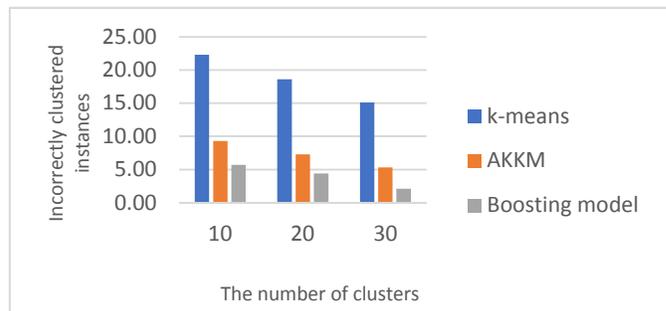


Fig. 6. Comparison of the quality of the algorithms (Iris benchmark)

According to the experimental results shown in Table 1 and in Figure 4, the developed boosting model can be more time-consuming than the popular AKKM method, but the working time can decrease significantly with an increase in the number of the objects for clustering. This can be explained by the probabilistic approach used in the model.

Table 2 and Figures 5 and 6 demonstrate high effectiveness of the developed boosting model in comparison to the classical k-means and AKKM methods. The comparative analysis of the quality of the solutions has shown that the quality of the solutions obtained with the boosting model is 40% higher than the analogs, which have less time complexity.

It should be mentioned that the results of the boosting algorithm work can vary depending on the algorithms used in the boosting model.

CONCLUSION.

The ongoing growth of the produced and transferred information creates the need for developing new algorithms for its processing. One of the main tasks in terms of data processing

problems is clustering. The paper presents the clustering task statement and the analysis of the popular modern and classical clustering algorithms, their benefits and shortcomings. The paper compares the boosting and bagging approaches and demonstrates their benefits and drawbacks.

The boosting procedures can be used for solving classification and clustering tasks. The authors developed a boosting model and a modified boosting algorithm to solve the clustering problem and compared them with the analogs in terms of time complexity and the percent of Incorrectly Clustered Instances (ICI).

The experiments were carried out with different numbers of the instances and datasets. The results show, that the boosting algorithm can be more time-consuming, but give a significant advantage in the quality of the solutions since the error of incorrectly clustered instances is 40% less on the average.

To improve the clustering quality and to reduce the time consumptions, we propose using different clustering algorithms for boosting. The time complexity can be optimized by using parallel paradigms of developing [23-26]. The quality of the algorithms included in the ensemble can change by adjusting the parameters of the probability matrix of including each element from the dataset into a cluster.

The developed boosting algorithm is suitable for solving semi-supervised learning tasks with a teacher. For example, we can use the modified matrix of including the element in a cluster for the classified elements and increase the probability of including into the cluster, where the element has been included. The other way is to reduce the task to clustering and to ignore classification.

Future work will be related to investigating the algorithms used for boosting based on bioinspired methods such as ant colony optimization [27], swarm optimization [28,29], and the hybrid algorithms based on the combining the principles of different bioinspired algorithms.

REFERENCES

[1] K. Wong, "A Short Survey on Data Clustering Algorithms", IEEE Second International Conference on Soft Computing and Machine Intelligence, 2015, pp. 64-68.

[2] R. Jacobson, *IBM Consumer products industry blog*. Industry insights, Apr. 2013. Accessed on: Apr. 05, 2019. [Online]. Available: <https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/>.

[3] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The Evolution of Boosting Algorithms – From Machine Learning to Statistical Modelling", *Methods Inf Med*, vol. 53, 2014, pp. 419 – 427.

[4] X. Donkuan and T. Yingjie, "A comprehensive survey of clustering algorithms", *Annals of Data Science*, vol. 2, no. 2, 2015, pp 165-193.

[5] A.A. Zaicev, V.V. Kureichik, and A.A. Polupanov, "Obzor evolucionnikh metodov optimizatsii na osnove roevogo", *Izvestiya YUFU. Tekhnicheskie nauki*, vol. 12 (113), 2010, pp. 7-12.

[6] V.V. Kureichik and Y.A. Kravchenko, "Bioinspired algorithm applied to solve the travelling salesman problem", *World Applied Sciences Journal*, vol. 22, no. 12, 2013, pp. 1789-1797.

[7] L.A. Gladkov, V.V. Kureichik, and Y.A. Kravchenko, "Evolutionary algorithm for extremal subsets comprehension in graphs", *World Applied*

Sciences Journal, vol. 27, no. 9, 2013, pp. 1212-1217.

V.V. Kurejchik, V.M. Kurejchik, and P.V. Sorokoletov, "Analiz i obzor modelej ehvolyucii", *Izvestiya Rossijskoj akademii nauk. Teoriya i sistemy upravleniya*, vol. 5, 2007, pp. 114-126.

[8] S.I. Rodzin and V.V. Kurejchik, "Sostoyanie, problemy i perspektivy razvitiya bioehvristik", *Programmnye sistemy i vychislitel'nye metody*, vol. 2, 2016, pp. 158-172.

[9] V.V. Kurejchik, V.V. Bova, and VI.VI. Kurejchik, "Kombinirovannyj poisk pri proektirovanii", *Obrazovatel'nye resursy i tekhnologii*, vol. 2 (5), 2014, pp. 90-94.

[10] V.V. Kurejchik and VI.VI. Kurejchik, "Bioinspirirovannyj poisk pri proektirovanii i upravlenii" *Izvestiya YUFU. Tekhnicheskie nauki*, vol. 11 (136), 2012, pp. 178-183.

[11] P.N. Druzhkov, N.Yu. Zolotyh, and A.N. Polovinkin, "Programmnaya realizaciya algoritma gradientnogo bustinga derev'ev reshenij", *Vestnik Nizhnegorodskogo universiteta im. N.I. Lobochevskogo*, vol. 1, 2011, pp. 193 – 200.

[12] A.J. Ferreira and M.A.T. Figueiredo, "Boosting Algorithms: a review of mehods, theory and applications", in: Zhang C., Ma Y. (eds) *Ensemble Machine Learning*. Springer, Boston, MA, 2012, pp. 35-85.

[13] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The evolution of boosting algorithms – From machine learning to statistical modeling", *Methods iInf Med*, vol. 53(6), 2014, p. 419 – 427.

[14] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm", in *Thirteenth International Conference on Machine Learning*, Bari, Italy, 1996, pp. 148–156.

[15] Y. Freund and R. Schapire, "A decision-theoretic generali zation of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, vol. 55(1), 1997, pp. 119–139.

[16] L. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms", Wiley, 2004.

[17] D. Martens, M. De Backer, R. Haesen, J. Vanthienen, M. Snoeck and B. Baesens, "Classification With Ant Colony Optimization," in *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 5, 2007, pp. 651-665.

[18] M.A.M. Shukran, Y.Y. Chung, W.C. Yeh, N. Wahid, and A.M.A. Zaidi, "Artificial Bee Colony based Data Mining Algorithms for Classification Tasks," *Mod. Appl. Sci*, vol. 5, 2011, pp. 217–231.

[19] Dr. Chandrika.J, Dr.B.Ramesh, Dr.K.R. Ananda kumar, and R.D. Cunha "Genetic Algorithm Based Hybrid Approach for Clustering Time Series Financial Data," *CSE*, 2014, pp. 39-52.

[20] E. Saraç and S. A. Özel, "Web page classification using firefly optimization," *2013 IEEE INISTA, Albena*, 2013, pp. 1-5.

[21] C. Radha, J. Rong, C.H. Timothy, and K.J. Anil, "Scalable Kernel Clustering: Approximate Kernel k-means", *Computer Vision and Pattern Recognition*, 2014.

[22] V.M. Kurejchik, V.V. Kurejchik, and S.I. Rodzin, "Modeli parallelizma ehvolyucionnyh vychislenij", "Vestnik Rostovskogo gosudarstvennogo universiteta putej soobshcheniya", vol. 3 (43), 2011, pp. 93-97.

[23] V.M. Kurejchik, V.V. Kurejchik, S.I. Rodzin, and L.A. Gladkov, "Osnovy teorii ehvolyucionnyh vychislenij", *Yuzhnyj federal'nyj universitet, Tekhnologicheskij institut. Rostov-na-Donu*, 2010.

[24] S.I. Rodzin, V.V. Kurejchik, "Teoreticheskie voprosy i sovremennye problemy razvitiya kognitivnyh bioinspirirovannyh algoritmov optimizacii", *Kibernetika i programirovanie*, vol. 3, 2017, pp. 51-79.

[25] V. Kureichik, D. Zaporozhets, and D. Zaruba, "Generation of bioinspired search procedures for optimization problems," *Application of Information and Communication Technologies, AICT 2016 - Conference Proceedings*, vol. 10, 2016.

[26] D. Martens, B. Baesens, and T. Fawcett, "Editorial survey: swarm intelligence for data mining," *Machine Learning*, vol. 82 (1), 2011, pp. 1-42.

[27] I.D. Falco, A. D. Cioppa, and E. Tarantino, "Evaluation of particle swarm optimization effectiveness in classification," *LNAI3849*, 2006, pp: 164- 171.

[28] M. Karnan, K. Thangavel, and P. Ezhilarasu, "Ant Colony Optimization and a New Particle Swarm Optimization algorithm for Classification of Microcalcifications in Mammograms," *16th International Conference on Advanced Computing and Communication*, 2008.

Modeling Technique of Large Signal Dynamics for Electromagnetic Levitation Melting System

Idan Sassonker
Dept. of Electrical and Computer Eng.
Ben-Gurion University of the Negev
 Beer-Sheva, Israel
 sassonke@post.bgu.ac.il

Moria Elkayam
Dept. of Electrical and Computer Eng.
Ben-Gurion University of the Negev
 Beer-Sheva, Israel
 moriael@post.bgu.ac.il

Alon Kuperman
Dept. of Electrical and Computer Eng.
Ben-Gurion University of the Negev
 Beer-Sheva, Israel
 alonk@bgu.ac.il

Abstract— The main advantage of Electromagnetic Levitation Melting system (ELM) is to solve the pollution problem in metal melting process. This paper provides an analysis of the system large-signal behavior for modeling the dynamics of ELM that can be used for control design. Moreover, the high frequencies utilized by wireless ELM have posed a challenge for simulation software in the form of long simulation times. In order to reduce simulation time, it is proposed to replace the analysis of the high-frequency-electrical-signals by only the envelope-signals and still get all the information on the system dynamics. A larger than 92% decrease in simulation time is observed. Experimental results are presented to validate the analysis and the simulation results.

Keywords— Magnetic levitation, envelope simulation, large signal analysis, resonant power converters, wireless power transfer

I. INTRODUCTION

Many melting systems in the industry include metal contact which leads to parasitic pollution. Electromagnetic levitation Melting System (EML) is a wireless, well-known technique used to avoid the risk of pollution. On electrically charged metallic object that moves across an electromagnetic field operates Lorentz force. The Lorentz force is vertical to the velocity vector of the metallic object and the magnetic field plane and its direction determined by the right-hand rule. This power can be used to levitate the metallic object. In addition, the eddy-currents induced in the metallic object by the time-varying electromagnetic field flows mainly at the "skin" of the conductor, between the outer surface and a level called the skin depth, δ . This fact leads to heat the object by Joule effect and when the temperature is higher than the melting temperature of the material, it will eventually melt it. This is known as the levitation melting process as described in [1] - [4]. In this paper it was chosen to melt a metallic ball. The melting system is as described in [6]. In the melting process, the electric power transfers wirelessly from the electrical system to the ball. As a result, the ball is able to levitate and melt. One of the things that requires attention in the design of an ELM is the system frequency, which affects the position of the ball and ultimately affects the levitation stability. This necessitates the use of frequency control to maintain system stabilization.

This paper provides an analysis of the system large-signal behavior for modeling the dynamics of ELM that can be used for control design. In addition, for simulation of high-frequency signal in PSIM software, height sample is required with reliability results. However, such sampling typically results in large memory consumption and a long-time simulation run relative to the real time analysis. In ELM simulation, it is enough to know the signals transmitted between the sub-electrical system and the sub-mechanical system behaviour in order to describe the floating ball dynamics. Apparently, these signals change slowly (around 3Hz), whereas a slower sample can provide the required information. The proposed methodology of envelope analysis, i.e. large signal modulation aimed to improve simulation time while maintaining full knowledge of the system dynamics available.

II. GENERAL SYSTEM

The forces acting on the object at levitation on the z-axis are shown in Fig. 1, where h is the height of the ball from the lowest current loop, and the current through the loops is sinusoidal with amplitude I_A and frequency f . Mg is the gravity force and $F_{Lorentz}$ is the Lorentz force [3] and [8] given by

$$F_{Lorentz} = F_s + F_d, \quad (1)$$

where F_s is the static component and F_d is the dynamic component of the force, modified as

$$F_s = I_A^2 \cdot k_s(h, f), \quad (2)$$

and

$$F_d = I_A^2 \cdot v \cdot k_d(h), \quad (3)$$

respectively, where v is the sphere velocity along the z-axis. $k_s(h, f)$ and $k_d(h)$ are function given in [6]. Their behavior is described in Figs. 2. Effects of the change in f on k_s can be neglected. Therefore, for determination of $f = f_{conv}$ as a frequency at steady-state, (1) can be written as

$$F_{Lorentz} = I_A^2 \cdot [k_s(f_{conv}, h) + v \cdot k_d(h)]. \quad (4)$$

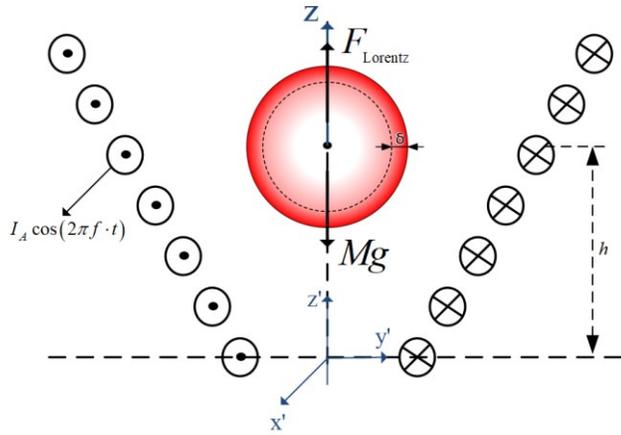


Fig. 1 An side cut showing the currents, forces acting on the object and the geometry.

The equivalent impedance of the load (ball) developed in [7], found that the inductance component, L_{ball} , and the resistance, R_{ball} , are functions of time, because at the dynamic movement of the sphere, h is function of time. Therefore, they can be written as $R_{ball}(t)$ and $L_{ball}(t)$. Fig. 4 shows the resistance and inductance as functions of height. As expected, with the increasing of the height, the equivalent inductance increases as well, as oppose to the equivalent resistance.

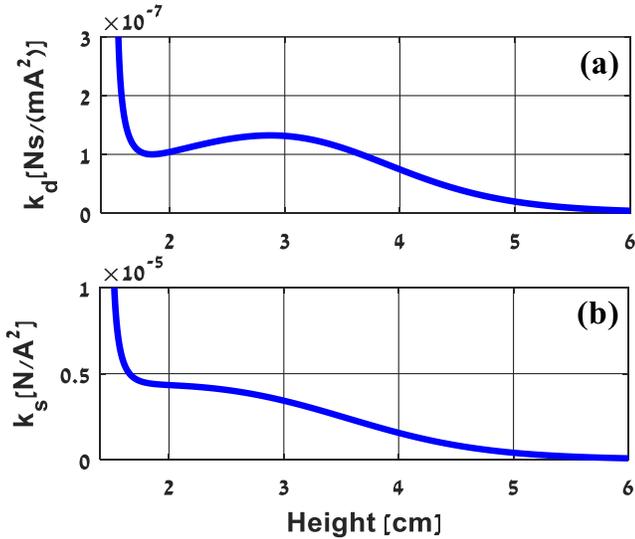


Fig. 2 (a) k_d function versus height and (b) k_s function versus height.

III. LARGE SIGNAL ANALYSIS

From (1), in order to calculate the Lorentz force, we should first analyze h, v_s and I_A dynamic behaviour at the ELM. In [6], it can be seen that the value of these signals changing much slowly than frequency switching. This fact gives the motivation to analyze only the envelope signals to get the all information on the dynamic movement of the ball. At resonance a simple equivalent circuit to the schematics of the system shown in [5] is given in Fig. 3. Where $v'(t)$ is the

first harmonic of output voltage from the inverter reflected to the load and $i(t)$ is the current through the coil. $L(t)$, $R(t)$ and C are the inductance, resistance and the capacitance reflected to the load, respectively, with

$$R(t) = R_{ball}(t) + R_{sys}, L(t) = L_{ball}(t) + L_{sys}, \quad (5)$$

where R_{sys} and L_{sys} are the load reflected parasitic resistance and inductance of the overall system. The behavior of the equivalent resistance and inductance of the ball presented in [6] are shown in Fig. 5.

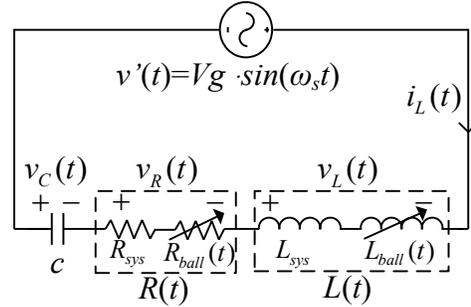


Fig. 3 A simple equivalent circuit to the system.

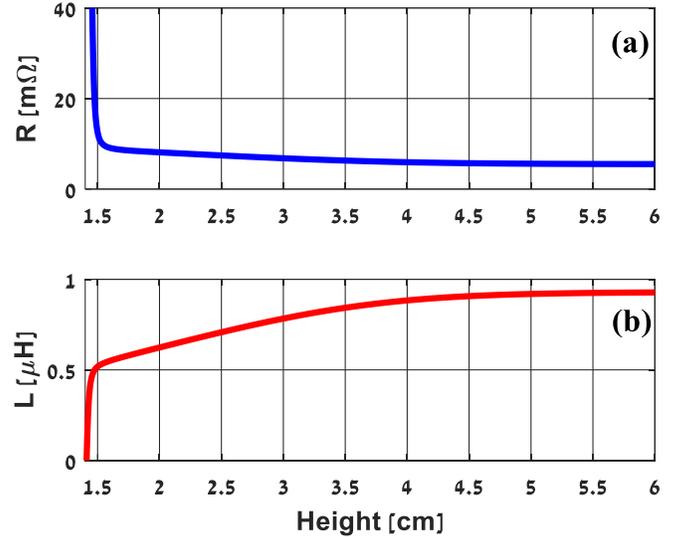


Fig. 4 The resistance (a) and the inductance (b) as functions of height.

and

$$V_g = (4/\pi) \cdot V'_{dc}, \quad (6)$$

where V'_{dc} is the load reflected voltage supply. Applying KVL leads to

$$v'(t) = v_R(t) + v_L(t) + v_C(t), \quad (7)$$

where the inductor voltage is given by

$$v_L(t) = \frac{d}{dt}(L(t) \cdot i(t)) = \frac{dL(t)}{dt} \cdot i(t) + \frac{di(t)}{dt} \cdot L(t), \quad (8)$$

and the resistor voltage is

$$v_R(t) = R(t) \cdot i(t), \quad (9)$$

and the circuit current can be described as function of the capacitor voltage as

$$i(t) = C \cdot \frac{d}{dt}(v_C(t)). \quad (10)$$

The current, $i_L(t)$, and the capacitor voltage, $v_c(t)$, can be approximated to be pure sinusoidal and can be written as the sum of sine terms and cosine terms [9]. Applying the above harmonic approximation in (7)–(10) gives

$$\begin{cases} i_L(t) = i_L^s \sin(\omega_s t) + i_L^c \cos(\omega_s t) \\ v_c(t) = v_c^s \sin(\omega_s t) + v_c^c \cos(\omega_s t) \end{cases}, \quad \omega_s = 2\pi\nu_s, \quad (11)$$

with ν_s is the frequency of the current. Their derivative is

$$\begin{aligned} \frac{di_L}{dt} &= \left[\frac{di_L^s}{dt} - \omega_s i_L^c \right] \sin(\omega_s t) + \left[\frac{di_L^c}{dt} + \omega_s i_L^s \right] \cos(\omega_s t) \\ \frac{dv_c}{dt} &= \left[\frac{dv_c^s}{dt} - \omega_s v_c^c \right] \sin(\omega_s t) + \left[\frac{dv_c^c}{dt} + \omega_s v_c^s \right] \cos(\omega_s t) \end{aligned} \quad (12)$$

Substituting (11) and (12) in (8) – (10), there is

$$\begin{aligned} \frac{di_L^s}{dt} &= \omega_s i_L^c + \frac{1}{L(t)} \left[V_g - \left(R(t) + \frac{d}{dt}(L(t)) \right) \cdot i_L^s - v_c^s \right] \\ \frac{di_L^c}{dt} &= -\omega_s i_L^s - \frac{1}{L(t)} \left[\left(R(t) + \frac{d}{dt}(L(t)) \right) \cdot i_L^c - v_c^c \right], \end{aligned} \quad (13)$$

and

$$\begin{aligned} \frac{d}{dt} v_c^s &= \frac{1}{C} I_s + \omega_s v_c^c \\ \frac{d}{dt} v_c^c &= \frac{1}{C} I_c - \omega_s v_c^s. \end{aligned} \quad (14)$$

The envelopes of the current through the coil and the capacitor voltage are calculated as

$$I_A = \sqrt{(i_L^s)^2 + (i_L^c)^2} \quad (15)$$

and

$$v_c = \sqrt{(v_c^s)^2 + (v_c^c)^2}, \quad (16)$$

respectively, where V_s^c is the sine component of the capacitor voltage and V_c^c his cosine component. The motion of the sphere in the vertical direction is given by Newton equation as

$$M \frac{d^2 h}{dt^2} = F_{\text{Lorentz}} \left(f_{\text{conv}}, h, \frac{dh}{dt}, I_A \right) - Mg. \quad (17)$$

Where M is the ball mass. The differential equations (13)–(16) represent system equations, also known as the large-signal model of the system and give information about h , ν_s and I_A dynamic behavior at the ELM.

IV. RESULTS

In order to verify the performance of the proposed algorithm, i.e. at low-frequency operation and compare it to the high frequency operation, simulation and experimental studies was carried out, based on the circuit in Fig. 5 with circuit parameters in Table I. The frequency of the switching signal is around 25 kHz. Simulation and experimental results of the current through the coil and capacitor voltage are shown in Fig. 6. It can be seen that simulation results following proposed modulation are seen to agree within practical results.

Fig. 7 shows the correlation of the coil current between the practical system (i_{Lp}), the high-frequency simulation (i_{Ls}) and its envelope in the low-frequency simulation (i_{Le}). Fig. 8 shows the capacitor voltage envelope (v_{ce}) on top of the full signal (v_{cs}). Experimental results closely match simulation outcomes, verifying excellent performance of the proposed large-signal modulation.

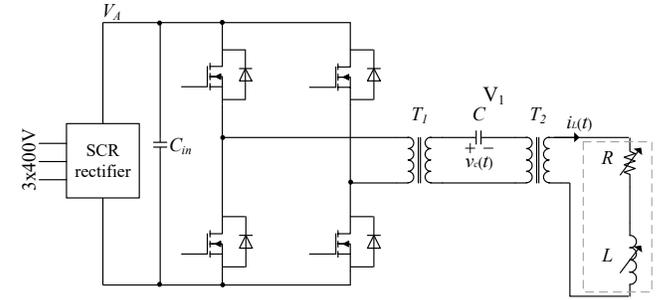


Fig. 5 A schematics of the system.

COMPONENTS VALUES

Symbol	Value	Units
T1	35:5	
T2	20:2	
Rsys	8.1	mΩ
Lsys	0.37	uH
M	22.7	gr
C	0.33	uF

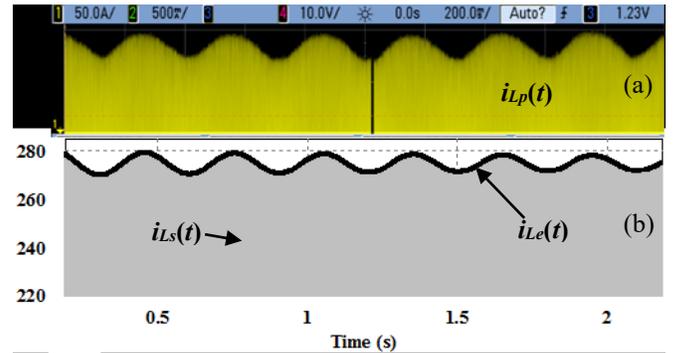
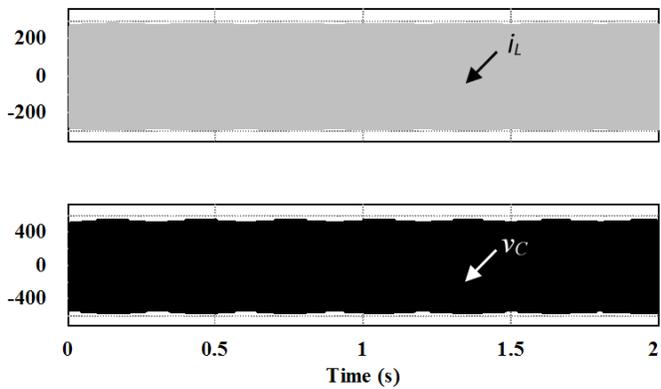
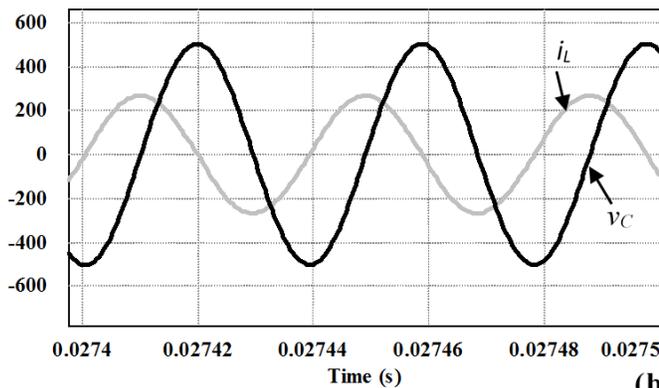


Fig. 7 The coil current – (a) Experimental results and (b) simulation results of high-frequency signal and its envelope.



(a)



(b)



Fig. 6 Experimental (left) and simulation (right) results of the current through the coil and voltage of the capacitor – full (a) and zoomed (b) signals.

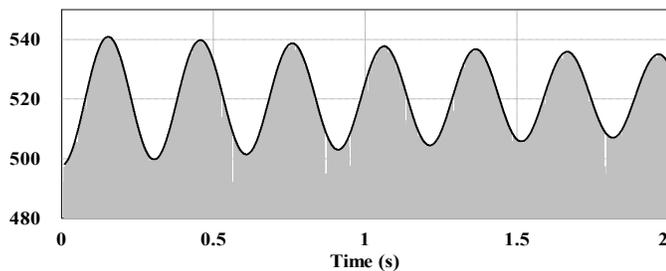


Fig. 8 Capacitor voltage envelope on top of the full signal – simulation results.

In addition, the velocity, height of the ball and the acceleration described the dynamic movement of the ball during the stabilization were simulated as shown in Fig. 9.

V. CONCLUSIONS

A simulation method of ELM using equivalent large-signal modulation was proposed herein. The proposed modeling technique allows us to simulate the system much faster by eliminating the high frequency components in the electrical system without data losses. Experimental results were also given to demonstrate and successfully verify findings presented in the paper.

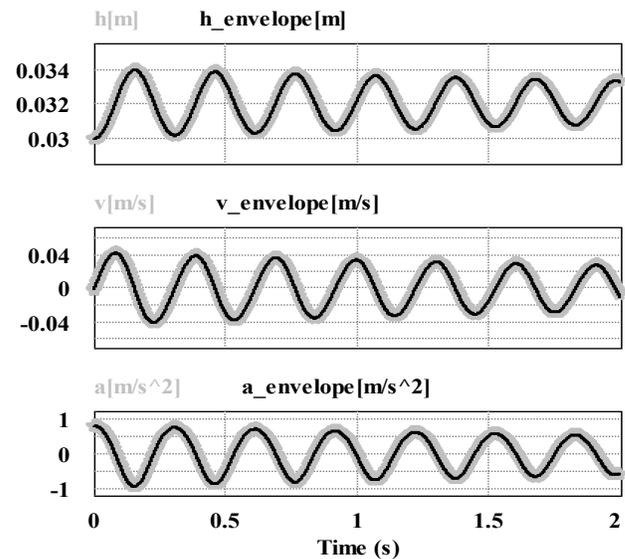


Fig. 9 The velocity, height of the ball from the lowest current loop and the acceleration – simulation results.

REFERENCES

- [1] E. C. Okress, D. M. Wroughton, G. Comenetz, P. H. Brace, and J. C. R. Kelly, "Electromagnetic Levitation of Solid and Molten Metals," *Journal of applied physics*, vol. 23, no. 5, pp. 545–552, May 1952.

- [2] W. Brisley and B. S. Thornton, "Electromagnetic levitation calculations for axially symmetric systems," *British Journal of Applied Physics*, vol. 14, no. 10, pp. 682–686, Oct. 1963.
- [3] B. O. Ciocirlan, D. G. Beale, and R. A. Overfelt, "Simulation of motion of an electromagnetically levitated sphere," *Journal of Sound and Vibration*, vol. 242, no. 4, pp. 559–575, May 2001.
- [4] S. R. Sagardia and R. S. Segsworth, "Electromagnetic levitation melting of large conduction loads," *IEEE Transactions on Industry Applications*, vol. IA-13, no. 1, pp. 49–52, Jan. 1977.
- [5] R. Rabinovici, V. Berdichevsky, M. Shvartsas and A. Shoihet. "Eddy-currents levitation system," in *IEEE 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, 2012.
- [6] I. Sassonker, M. Shvartsas, A. Shoihet and A. Kuperman, "Modeling of Electromagnetic Levitation Melting System with Experimental Validation", in *Proc. 2018 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Eilat, 2018.
- [7] G. Lohöfer, "Magnetization and impedance of an inductively coupled metal sphere," *Int. J. Eng. Sci.*, vol. 32, (1), pp. 107-117, 1994.
- [8] William R. Smythe, *Static and Dynamic Electricity*. Hemisphere Publishing Corporation, 1989.
- [9] Z. U. Zahid *et al.*, "Modeling and Control of Series–Series Compensated Inductive Power Transfer System," in *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 3, no. 1, pp. 111-123, March 2015.

On a Method for Segmentation of Memory Instances with Row Redundancies

Karen Amirkhanyan
ET&R department
Solutions Group
Yerevan, Armenia
kamirkha@synopsys.com

Valery Vardanian
ET&R department
Solutions Group
Yerevan, Armenia
vvardani@synopsys.com

Abstract—In this extended abstract, we proposed a method for “segmentation” of large memory instances with global redundant rows into memory segments with local redundant sub-rows allowing, thus, to split the memory into segments and the redundant rows - into local redundant sub-rows. A segment has its own local redundant sub-rows. Each redundant sub-row is a corresponding segment of a global redundant row. Thus, we increase significantly the number of redundant sub-rows and repair the faults/defects occurring in a memory segment with its local redundant sub-rows. For large memory instances, this approach will allow to increase the repair coverage of the memory instance with negligible hardware and time overheads. In general, a trade-off between fault/defect coverage and hardware/time overheads should be done.

Keywords—Memory system, repair, local/global redundancy, repair coverage.

I. INTRODUCTION

Many approaches were proposed earlier (see [1]-[8]), such as “shared BISR”, grouping, reusing, shared functional/test buses, shared redundancies/registers etc. that improve the overall repair coverage of a memory instance.

Today’s complex SoC usually contains hundreds and even thousands of memory instances. Synopsys’s DesignWare Multi-Memory Bus (MMB) [1], based on ARM’s shared test bus, also used the test bus for efficient test of multiple memory instances (known as a Memory System (MS)) attached to the bus. Mentor Graphics [9] proposed to use a functional bus for its efficient testing in addition to the test bus. It was suggested in [9] to use the functional bus for efficient repair of memories via “redundancy sharing” due to which time and hardware are saved during the repair, and the repair coverage is improved. Redundancy sharing that was proposed in [10] means the following. If a defect/fault is detected in a memory instance of the MS, and there is a free (unused) redundancy available elsewhere, possibly in another memory instance within the same group of homogeneous instances of the MS, then it can be used for repair bringing to a significant reduction of the area.

The redundant elements remaining after manufacturing repair can be used in the field during test & repair sessions, as well as soft repair, performed periodically after power-up (see [11]). A reparable memory instance with redundancies used its own (local) redundancies. It is based on the idea of “a sharing mechanism for redundancies” within the same group of homogeneous memory instances in the MS. The BIST process

is performed by means of the shared test bus, and the repair is performed by means of the functional bus (see [10]).

In this paper, we propose another method for improving repair coverage and yield of very large (Number of Bits in the word of the memory is >128 bits) memory instances in SoCs. Since a reparable memory instance usually has very limited resources (usually, a few rows of redundant rows for repair purposes (we consider the case of availability of redundant rows only) there may occur situations when the redundancies are not enough to repair the memory instance. Let us suppose that we detected in our memory instance with k redundant rows more than or equal to $k+1$ faults/defects such that any two of them are not in the same row. Then since only row redundancies are assumed to be available, and for each fault we need one redundant row to repair a fault/defect, there will be a fault/defect that could not be repaired.

The number of faults/defects in a memory instance is determined by multiplying the area of the instance by the defect density of the memory instance that shows the number of faults/defects in a unit area of the memory. Usually, the defect density for each type of memory is known beforehand. Thus, the number of predicted faults/defects is determined by the area of a memory instance. If the number of predicted faults/defects to be repaired is too high exceeding the number of available redundant rows, then, in the worst case, it is possible that the memory will not be repairable in case when no any two faults/defects are in the same row. Thus, each fault/defect will require a new redundant row to repair it. Thus, to decrease the number of faults contained in a memory unit we need to decrease its area. Therefore, the idea of segmented memories emerges. We need to split the memory into units (let us name them “segments”) that possess smaller area, and, according to the known formula for the number of predicted faults/defects in a memory area (segment), does not exceed the number of available local redundant rows for the segment so, in the worst case, when one redundant row is needed for each fault/defect to repair it, we will be able to repair the memory segment for sure. Thus, our main idea is to split the memory into memory segments such that they are all reparable by their local redundant rows in the worst case. Each segment should have at least so much redundant rows as the predicted number of faults/defects in the segment. The memory instance is split into homogeneous segments of the same size s to be determined later. All segments have the same area and defect density. Hence the number of predicted faults/defects in all segments is the same. Thus, the number of the predicted

faults/defects in a segment equals to the number of all faults/defects in our initial memory instance divided by the number of segments the memory is split into. Note that we consider the case when the faults/defects are distributed over the area of the memory with equal probability.

A similar methodology is being currently developed for large repairable memory instances with column and both row and column redundancies. Due to the page limitations and for the sake of simplicity, we will constraint ourselves in this extended abstract with consideration of a simple case when the memory instance has only row redundancies.

II. SEGMENTATION OF A MEMORY INSTANCE

Memory Instance with k redundant rows (no segments)

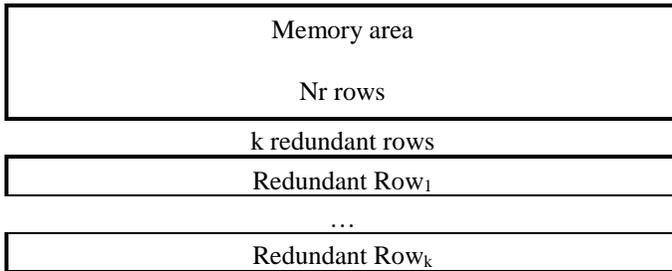


Fig. 1. Memory instance with k redundant Rows

Consider a memory instance (see Figure 1) with N_r rows and k redundant rows. Note that for the sake of simplicity we consider here only the case of availability of redundant rows only. The case for redundant columns or both rows and columns will be considered in the future. Note that the repair coverage depends on the location of faults/defects detected by BIST. In the worst case when $k+1$ faults/defects were detected by BIST when no any two faults/defects were in the same row then each fault/defect would require a redundant row for repair. If the memory and correspondingly the k redundant rows are split into s segments (see Figure 2) then the segmented memory will have $s \cdot k$ redundant sub-rows. Each sub-row in a certain segment can be used to repair a fault/defect detected in that certain segment. Thus, segmentation of a memory instance will increase the number of redundant elements and, as a result, the repair coverage of the segmented memory will increase drastically. Note also that increasing of the number of redundant sub-rows will increase also the hardware overhead connected with additional control logic, number of redundancy registers, etc. In Section V, we will estimate the hardware overhead required by the process of segmentation.

In this extended abstract, we proposed a method that splits a given big memory instance M into s “segments”. By a segment we mean a memory region where a given number of columns are included with complete bit-lines without any interruption in the segment. The recommended number of segments s is to be determined later in Sections IV and V. Figure 2 depicts an example of segmentation of a memory instance M into s segments of the same size.

Memory Instance M with s segments

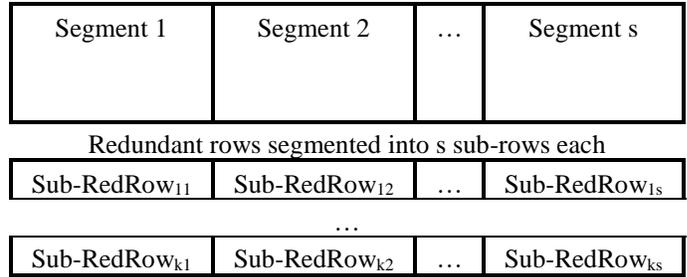


Fig. 2. Segmented Memory instance with segments and redundant Sub-Rows

In general, note that this requirement is not necessary, i.e. the segments may be of unequal size. In this extended abstract, we consider only the simple case when all segments are of the same size. If the memory instance had initially k redundant rows, $k \geq 1$, then note that each redundant row is also split into s redundant sub-rows. Hence each memory segment (denote the i -th segment S_i) will be assumed repairable with k redundant sub-rows. Thus, after modifications our memory instance should have $k \cdot s$ local redundant sub-rows, with k local redundant sub-rows for each segment. Note that if during a BIST session $k+1$ faults/defects were detected in a certain segment then the memory instance will not be repairable if no any 2 faults/defects would be found in the same row. Note, however, the probability of this event must be very low, in other words, almost impossible.

Note that formerly, if a fault/defect was detected in the former memory instance then for its repair we used a complete redundant row R_t , $1 \leq t \leq k$, of the same size as the instance. Now, for the segmented memory, when it is split (segmented) into, say, s segments, the redundant row R_t – into s segments of local redundant sub-rows R_{t1}, \dots, R_{ts} , and the fault/defect is detected in a certain segment S_j , $1 \leq j \leq s$, then if we still have an available redundant sub-row then one of the available sub-rows can be used for the repair of the fault/defect. Note that:

- segmentation of redundant rows into sub-rows brings to more efficient usage of redundant resources,
- increases the number of redundant units. Instead of k redundant rows there will be available $k \cdot s$ local redundant sub-rows after the segmentation process. Since in nowadays technology memories the number of faults/defects is very few then after manufacturing test and repair most of the remaining redundant sub-rows can be used for further repair in the field.

III. SOME DETAILS OF IMPLEMENTATION

A. Conventional Hardware Implementation

The memory structure was presented in detail [10], [11], but in this extended abstract we will highlight only implementation of the decoder of rows in SRAM memories.

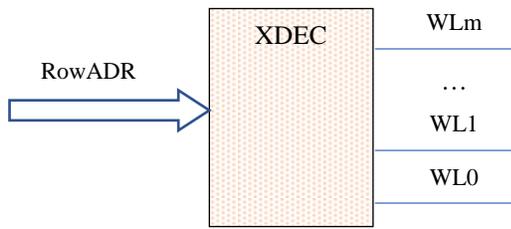


Fig 3. XDEC functional view

In the former publications [11], [12], [17], [18], the decoder of rows usually was mentioned as abbreviation XDEC because it included the memory banks decoders and the timing formation blocks for row signals as well. Based on the logical address, XDEC decodes and forms the row select signal for the corresponding row of the memory bit cells (see Figure 3).

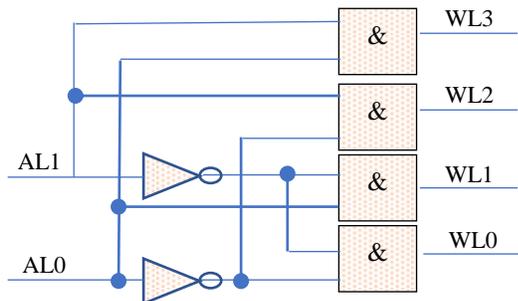


Fig 4. 2x4 regular decoder

By using the logical input Row Address (denoted RowADR in Figure 3) the row decoder decodes and activates the corresponding word line selection signals (WLM), m is the number of the rows in the memory instance. In Figure 4, a simplified example of the regular decoder scheme (2 x 4) is presented for rows. In general case, the decoding has the scrambled structure [11], [12], [17]. In the case presented in Figure 4 that scrambling is the regular one and the Truth table matches with that of the given in the Table 1.

Table 1. The Truth table of the Regular 2x4 decoder

State of the Word Line				State of the Address Lines	
WL3	WL2	WL1	WL0	AL1	AL0
0	0	0	1	0	0
0	0	1	0	0	1
0	1	0	0	1	0
1	0	0	0	1	1

WL0 is activated when the address lines AL0 and AL1 obtain 0 values.

WL1 is activated for the address AL1=0, AL0=1;

WL2 - for the case AL1=1, AL0=0;

WL3 will be activated for the address AL1=1 and AL0=1.

B. Implementation of the Proposed Mechanism of Segmentation

In the conventional memory WLM activates the memory bit cells of the whole word line, but in the Segmented memory the word line is divided into s parts, as result of segmentation. In the Segmented memory we must manage the divided word line.

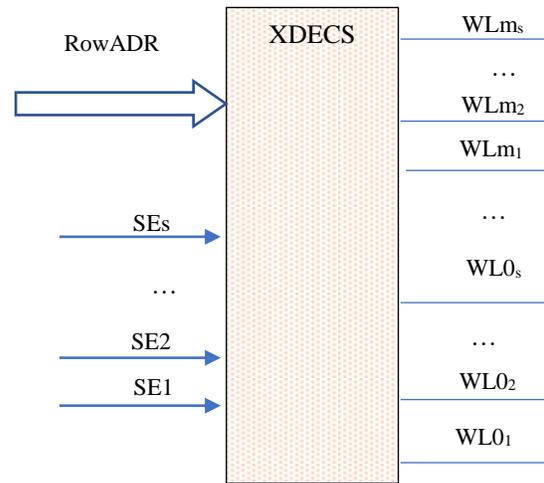


Fig 5. XDECS functional view

All parts of the segments are independent and should work with any combination. In the normal functional mode, all the segments work simultaneously (as in the conventional memory). In the case if any number of faults/defects will be detected in the cells of a memory row we must deactivate any segment in the word with a fault/defect and replace it (repair) by an available redundant sub-row corresponding to the segment. Let us suppose that the number of segments in the memory is "s" then each word line of the memory should be

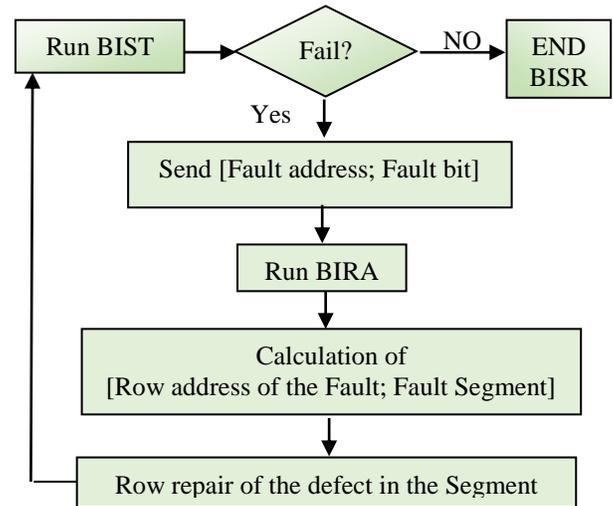


Fig 7. BISR algorithm for the Segmented memory

controlled by "s" number of WL signals. In Figure 5, an example of the functional view of XDECS ("row and segment" decoders) with the segmented WL is presented. The Segment Enable (SE) signals together with WLM signals form the word line control signals (see Figure 6). The presented control of the outputs of the row decoder gives the possibility to repair a defect in the segment using the corresponding redundant sub-

row (a local redundant sub-row belonging to the corresponding segment obtained from a redundant row of M segmented into s sub-rows, one sub-row for each redundant row).

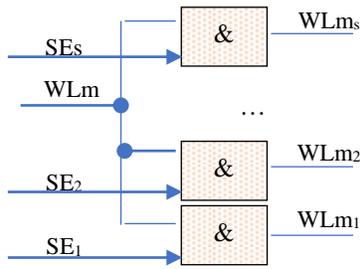


Fig 6. Segmented Word-Lines

This method increases the number of redundant elements in the segmented memory and evidently improves the repair coverage of the segmented memory instance. If formerly the memory instance M had, say, k redundant rows then after segmentation of M and its redundant rows we have sk redundant sub-rows. Each segment of M now will have k local redundant sub-rows that, if necessary, will be used for the repair of the corresponding segment if a fault/defect will be detected in the segment. Since in nowadays technologies, as a rule, the number of defects/faults is very low the number of local redundant sub-rows $s \cdot k$ will be greater enough than the actual number of faults/defects detected during the BIST session many redundant sub-rows may be left for future repair during test & repair sessions applied periodically in the field.

In Figure 7, the flowchart of the modified BISR algorithm is presented for the segmented memories. As for the conventional SRAM instance, the first step of the memory testing is the BIST run [14], [15] for checking the possible availability of faults/defects in the memory instance. For the embedded SRAM memories, the recommended BIST algorithm is to use a certain memory test algorithm for detection of faults/defects in SRAM memories. A March test algorithm [11], is of linear complexity developed for detection of a certain class of faults/defects in the memory. Earlier many efficient (minimal) March test algorithms were developed for detection, diagnosis or localization of certain classes of faults/defects in memories. If the test passes successfully then it means that the memory is free from the considered class of faults/defects. If BIST fails (a fault/defect is detected) in the memory area the BIST returns the following information about the defect: the logical address of the fault/defect and the faulty/defective bit. The next step is the run of the Built-In Redundancy Allocation (BIRA) algorithm. Based on that input of the logical address of the fault/defect and the scramble information of the memory instance [14] BIRA calculates the physical Row address of the fault/defect and the corresponding segment with the fault/defect for the execution of row repair of the fault/defect. And the last step is the rerunning of BIST for assuring the memory is free of a fault/defect.

IV. ESTIMATION OF THE NUMBER OF SEGMENTS

Introduce the following notations:

D_M - defect density of an SRAM memory M ,

A_M - area of a memory instance M with redundancies, Then it is predicted (see [10], [16]) that the number of possible defects F_M in a memory instance M with redundancies can be determined according to the formula below:

$$F_M = D_M \cdot A_M \cdot (1)$$

Let R_M be the number of redundant Rows available in M . Evidently, if

$$R_M \geq F_M \quad (2)$$

then all the faults that are predicted to occur in M can be repaired since in the worst case each fault/defect will require a redundant row, and due to inequality (2) it will be possible to repair all faults/defects in M .

Now, suppose, $R_M < F_M$. In this case, we propose to split the memory instance M into s segments. Note that we consider the case when the faults/defects are distributed over the whole area of M with equiprobable outcome. In this case, the number of predicted faults/defects in every segment of M will be the same. Since all s segments of M are of the same area then A_M/s will be the area of a certain segment. Since the segments should all have the same defect density and number of faults/defects then each segment will predictably contain F_M/s faults/defects. At the same time, if the memory instance M had k redundant rows then each segment should have k redundant local sub-rows since from each redundant global row in M we extract a certain local redundant sub-row. Thus, each segment will be repairable with k redundant sub-rows if and only if $k \geq F_M/s$. From this inequality we obtain the following condition for s , the number of segments the memory instance is recommended to split into:

$$s \geq F_M / k. \quad (3)$$

From formulas (1) and (3) we obtain the following main inequality for the number of segments to define:

$$s \geq D_M \cdot A_M / k.$$

V. ESTIMATION OF THE AREA OVERHEAD

It is obvious that the approach proposed in this paper brings to some area and time overheads in the segmented memory instance. If the initial memory instance had k redundant rows, and the memory instance was split into s segments, then the segmented memory will have $s \cdot k$ redundant sub-rows, each segment having k redundant sub-rows. Thus, the increase of redundant sub-rows, k redundant sub-rows for each segment, will obviously increase the repair coverage of the memory instance. However, since each redundancy requires a register for storing the address of a fault/defect in a certain segment to repair then $s \cdot k$ redundancy registers should be available in the segmented memory. The structure of the redundancy registers is considered the same in the initial and the segmented memory instances. The latter consists of two components. The first is the following: increasing of the number of redundant elements assumes increasing of the redundant control blocks such as the redundancy registers containing the corresponding fault/defect address and their control logic elements (see [13]) As many times as the number of redundancy registers will be increased, correspondingly the amount of the hardware of the

control blocks will be increased as well. For the redundant control logic added, the area overhead (S_L) will be equal to

$$S_L = S_R \cdot (s \cdot k - k) = S_R \cdot (s - 1) \cdot k$$

where S_R is the total area of the logic elements used for implementation of the control block for one redundancy, in this case – redundant row.

The second component of the area overhead is the control logic which must be implemented for the control of the segments in each row of the memory area of the instance. Deactivation of the redundant sub-rows in a redundant row, being split into s redundant sub-rows, is implemented by the additional control logic (see Figures 5 and 6) to activate the corresponding redundant sub-row corresponding to a fault/defect detected by BIST and localized in a certain segment. For this parameter, the area overhead (S_F) will be calculated by the following way

$$S_F = s \cdot S_{AND} \cdot N_r,$$

where S_{AND} is the area of the logical AND element of the used technology of the memory instance, and N_r is the number of rows in the memory area of the instance. The total area overhead for the instance M (denoted S_M) will be equal to:

$$S_M = S_L + S_F.$$

Finally, we calculate the percentage of the overhead:

$$S_{\%} = ((S_M - S_I) / S_I) \cdot 100\% = (S_M / S_I - 1) \cdot 100\%$$

where $S_{\%}$ is the overhead in percentage, S_M is the instance area of the segmented memory and S_I is the instance area of the initial non-segmented memory.

We can see from the formulas for the hardware overhead for a memory instance that consists of s segments it has linear dependence on the number of the segments. We recommend using this method by doing a careful trade-off between the number of segments that allows using redundancies more efficiently and increasing the repair coverage of the memory instance tending to minimize s , the number of the segments, versus the hardware overhead for additional control blocks. Seemingly this approach will work good for initially big memory instances with bigger amount of bit-cell area.

VI. CONCLUSION

In this paper, we proposed a “memory segmentation mechanism” for the repair of faults/defects in large memory instances with limited number of redundant rows. To improve the repair coverage of a large memory instance we need to increase the number of redundant resources. We suggest splitting the memory instance into several number s of “memory segments” with their local redundant sub-rows so that all the defects detected in each memory segment will be repaired by their local redundant sub-rows. The number of faults/defects predicted to occur in a segment is determined by multiplication of the memory defect density with the area of a segment. The number of segments the memory instance is to be split into is determined in such a way that the number of faults/defects in each segment should not exceed the predicted number of faults/defects in a segment.

In this paper, for the sake of simplicity, we considered only the case of availability of row redundancies only to repair the

memory instance. In the future, we are planning to extend this research for the Memory Systems and the cases when both redundant columns and rows are available in the reparable memory instance under consideration.

REFERENCES

- [1] Synopsys Press Release: “Imagination Technologies Adopts Synopsys STAR Memory System for Embedded Memory Test and Repair for New MIPS Processor,” <http://news.synopsys.com/2016-11-15-Imagination-Technologies-Adopts-Synopsys-STAR-Memory-System-for-Embedded-Memory-Test-and-Repair-for-New-MIPS-Processor>.
- [2] C.-L. Su, R.-F. Huang, and C.-W. Wu, “A processor-based built-in self-repair design for embedded memories,” in *Proc. 12th Asian Test Symp.* pp. 366-371, 2003.
- [3] T.-W. Tseng, J.-F. Li, and C.-C. Hsu, “ReBISR: A reconfigurable built-in self-repair scheme for random access memories in SoCs,” in *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 18, pp. 921-932, June 2010.
- [4] C.-D. Huang, J.-F. Li, and T.-W. Tseng, “ProTaR: An infrastructure IP for repairing RAMs in System-on-Chips,” in *IEEE Trans. Very Large Scale Integration Systems*, vol. 15, pp. 1135-1143, Oct. 2007.
- [5] T.-W. Tseng, J.-F. Li, and C.-S. Hou, “A built-in method to repair SoC RAMs in parallel,” in *IEEE Design & Test of Computers*, vol. 27, pp. 46-57, November-December 2010.
- [6] C.-L. Su, R.-F. Huang, and C.-W. Wu, “A processor-based built-in self-repair design for embedded memories,” in *Proc. 12th Asian Test Symposium (ATS'03)*, pp. 366-371, 2003.
- [7] S.-K. Lu, Z.-Yu Wang, Yi-Ming Tsai, and Jiann-Liang Chen, “Efficient Built-In Self-Repair Techniques for Multiple Repairable Embedded RAMs”, *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 4, pp. 620-629, 2012.
- [8] G. Wang, C. Chang, “Design and Implementation of Shared BISR for RAMs: A Case Study,” *IEEE Autotestcon, CA*, pp. 1-7, 2016.
- [9] D. Sargent, “Viewpoint: memory BIST for shared-bus applications”, *EDN network*, February 16, 2012, <http://www.edn.com/design/test-and-measurement/4389500/Viewpoint-Memory-BIST-for-shared-bus-applications>. Shyue-Kung Lu, Hao-Cheng Jheng, Hao-Wei Lin, Masaki Hashizume, and Seiji Kajihar, “Built-in Scrambling Analysis for Yield Enhancement of Embedded Memories”, *IEEE 23rd Asian Test Symposium*, pp. 137-142, 2014.
- [10] K. Amirkhanyan, S. Shoukourian, and V. Vardanian, “Design of reparable memory systems with shared row redundancies”, in *IEEE 11th Int'l Conf. “Computer science and information technologies” (CSIT'2017)*, Yerevan, pp. 144-147, 2017.
- [11] A.J. van de Goor “Testing Semiconductor Memories: Theory & Practice”, ComTex Publishing, 1998, 512P.
- [12] A.J. van de Goor, “Address and Data Scrambling: Causes and Impact on Memory Tests”, *DELTA*, pp. 128-136, 2002.
- [13] A. Bosio, L. Dilillo, P. Girard, S. Pravossoudovitch, A. Virazel, “Advanced Test Methods for SRAMs. Effective Solution for Dynamic Fault Detection in Nanoscaled Technologies”, *Springer* 2010, - 171 p.
- [14] Y. Zorian and S. Shoukourian, “Embedded-memory test and repair: Infrastructure IP for SoC yield,” in *IEEE Design & Test of Computers*, vol. 20, pp. 58-66, May-June 2003.
- [15] K. Darbinyan, G. Harutyunyan, S. Shoukourian, V. Vardanian, and Y. Zorian, “A robust solution for embedded memory test and repair”, in *Proc. IEEE Asian Test Symposium*, pp. 461-462, 2011.
- [16] J.A. Cunningham, “The use and evaluation of yield models in integrated circuit manufacturing”, *IEEE Trans. On Semiconductor Manufacturing*, vol. 3, No. 2, pp. 60-71, May 1990.
- [17] Alexanyan K., Amirkhanyan K., Shoukourian S., Shubat A., Vardanian V., Zorian Y., “Various methods and apparatuses for memory modeling using a structural primitive verification for memory compilers”, *US Patent No. 8,112,730*, pp. 1-22, 2012.
- [18] Aleksanyan K., Amirkhanyan K., Shoukourian S., Vardanian V., Zorian Y., “Memory Modeling Using an Intermediate Level Structural Description”, *US Patent, No 7768840*, pp. 1-17, 2010.

Elaboration of the Functioning Algorithm of Three – Dimensional Model of Computer System Safety

Victor V. Zhilin

Information System Cybersecurity Chair
Don State Technical University
Rostov–on–Don, Russia
zhilin95@inbox.ru

Irina I. Drozdova

Information System Cybersecurity Chair
Don State Technical University
Rostov–on–Don, Russia
irina_23011995@mail.ru

Ivan A. Sakharov

Information System Cybersecurity Chair
Don State Technical University
Rostov–on–Don, Russia
sakharov.i.a@yandex.ru

Larissa V. Cherckesova

Mathematics and Computer Sciences Chair
Don State Technical University
Rostov–on–Don, Russia
chia2002@inbox.ru
ORCID 0000–0002–9392–3140

Vitaliy M. Porksheyev

Applied Mathematics Chair
Rostov–on–Don, Russia
spu–46@donstu.ru

Olga A. Safaryan

Information System Cybersecurity Chair
Don State Technical University
Rostov–on–Don, Russia
safari_2006@mail.ru

Andrey G. Lobodenko

Information Systems and Radioengineering
Don State Technical University
Rostov–on–Don, Russia
andrey@sssu.ru

Sergey A. Morozov

Information Systems and Radioengineering
Don State Technical University
Rostov–on–Don, Russia
andrey@sssu.ru

Abstract—in this report, the known models of safety were considered. The general description of algorithm of three-dimensional model of safety of computer systems was provided. The relations arising between subjects of model of safety were described, and variations of use of temporal parameter are considered. Besides, the main shortcomings and possible threats of safety three-dimensional model of were revealed.

Keywords—safety model, computer system, object, subject, access, time, array, powers, operations, database, hierarchical relations.

I. INTRODUCTION

Currently, one of the actual problems of the theory of computer security is the development of mathematical models of security access control and information flow in modern computer systems (CS) [1].

This problem arises both in the theoretical analysis of CS safety with the use of their formal models, and in the testing of COP protection mechanisms with the use of procedures, methods and tools of automation and computer simulation [2].

The purpose of this paper is to develop an algorithm for the functioning of a new security model that combines the advantages of the currently known models.

It is possible to allocate the following tasks [3]:

1. Consideration of known safety models.
2. The allocation of the advantages and disadvantages of models.
3. Description of own development.
4. Detection of vulnerabilities of characteristic development.

For a start it should be noted what the so-called model of safety is necessary for. Its purpose is to formulate the safety requirements that the system should possess. The security model analyzes the properties of the system and determines the flows of information that are allowed in it [4].

By consideration of safety model such concepts as access to information; rules of differentiation of access, an object and the subject of access are used. Let us give precise definitions of each of them. Access to information implies familiarization with the information and carrying out operations such as processing, copying, modification and destruction of information [5].

Access control rules – a set of rules governing the access rights of subjects to access objects [6]. The object of access is a unit of information resource of the automated system, access to which is regulated by the rules of access control [7].

An access subject is a person or a process whose actions are regulated by the access control rules [8].

II. BASE METHODS

The security models themselves are distinguished by their security policies. Security policy may depend on the specific technology of information processing, technical and software tools used, and the location of the organization in which the security policy is described [9].

Let us consider the main models of safety from which it is possible to distinguish, five-measuring space Hartson's, model based on access matrix and the take model–Grant [10].

Five-measuring Hartson's space of received the name from quantity of the main sets:

1. Established authority (A).
2. Users (U).
3. Operations (E).
4. Resources (R).
5. States (S).

Thus, the security scope of this model will be a Cartesian product of these sets. In this case, the access will be considered requests entered by users to perform any operations on the system resources [11].

Users request access to resources. If they succeed, the system enters a new state. The query in this case looks like a four-dimensional tuple of the form $q = (u, e, R', s)$, where $u \in U$, $e \in E$, $s \in S$, $R' \subseteq R$ (R' –the requested resource set). The advantage of this model is that you can control access to an individual operation on a single object.

It should be noted that, due to the time-consuming algorithm, Harrison's security model has not been widely applied, unlike the model based on the access matrix. It is a rectangular table with rows corresponding to access subjects and columns corresponding to access objects [12].

The cells in such Table I describe all operations on objects that are allowed to the subject.

TABLE I. ACCESS MATRIX

	O ₁	O ₂	...	O _j	...	O _N
S ₁		w				
S ₂	r					
...						
S _i				r, w		
...						
S _M						e

In the table under the *w* refers to the recording of the object, *r* is the reading of the object under *e* – start. The values recorded in the table cells determine the types of safe access of the corresponding subject to the corresponding object [13].

This method of displaying access rights is lot more convenient in contrast to five-dimensional space of electronic discovery.

The disadvantage is that access rights exist separately from data. Nothing prevents a user who has access to classified information, write it to a file accessible to all, or replace it with a useful utility "Trojan" analogue [14].

As a basis, the Take-Grant model uses the notion of graphs [14]. Recall that a graph is an object that contains vertices and edges. Either objects or subjects are used as nodes in this model [15]. These nodes (vertices) are connected by arcs (edges). The values of these arcs characterize the rights that such a node has. There are four different conversion rules: *take*, *grant*, *create*, and *remove*.

Taken rule allows a subject to take the rights of another object, grant allows the subject to grant its own rights to another object, create allows the subject to create new objects, and remove removes the rights of subject that it has over an object.

For convenience, we introduce the following notation:

O – set of objects.

S – lot of subjects.

R = {r₁, r₂, r₃, r₄, ..., r_n} ∅ {t, g} - set of access rights.

t – the right to «take» access rights.

g – right to «grant» access rights.

G = (S, O, E) is a finite, labeled, oriented loop-free graph.

× – objects, elements of set O.

• – subjects, the elements of the set S.

Figure 1 shows all the rights of this model. Note that these rules generally look as follows: take(r, x, y, s), grant(r, x, y, s), create(r, x, s), remove (r, x, s).

Thus $r \in R, s \in S, x, y \in O$ the vertices of the graph G.

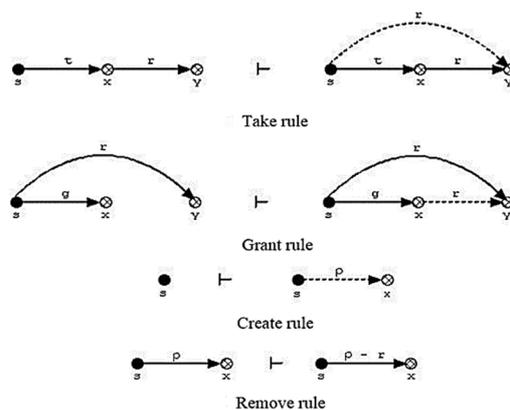


Fig. 1. Conversion rules.

Use model of safety of the computer take-grant systems can be established how the condition of system at change of the rights of subjects over objects changes.

The take-grant model is intended for the analysis of systems of protection with discretionary security policy. It describes the rules under which the transfer or unauthorized receipt of access rights. From a practical point of view, basically, there are quite simple relationships of objects [16]. The take and grant rules themselves are used quite rarely. Read and write permissions are most commonly used.

Thus, discretionary access control, which contains two basic rules, can be called the basis of discretionary security policy [13]:

- all subjects and objects used in this model must be uniquely identified or identified;
- the access rights of the system subject to the object are determined from some rule that is not described in advance.

The advantage of discretionary security policy is a simple implementation of protection mechanisms, as modern automated systems meet the rules of a specific security policy.

The disadvantage is the lack of flexibility in configuring the system. Besides, when using discretionary policy there is a question of what rules of distribution of access rights should be used and as they influence safety of system in general.

III. THREE-DIMENSIONAL MODEL OF COMPUTER SYSTEMS SAFETY

General description of the model

As you can see, each of the models considered earlier has its advantages and disadvantages. Based on these factors, a model based on three parameters – subject, object, time – will be presented below. In the future, such a three-dimensional model will be called a SOT – array.

The main difference between this model and the previously considered is the presence of another element – time. When determining the subject's right to perform any operation on the object, the value stored in the three-dimensional array is taken into account.

A two-dimensional matrix is a rectangular table with rows corresponding to access subjects and columns corresponding to access objects. Cells in such a table describe all operations on objects that are allowed to the subject.

In addition to subjects and objects, the three-dimensional array uses a time parameter, which is also taken into account when determining the allowed operations on the object. The

list of allowed operations is specified in the corresponding coordinates of this SOT array. It should be noted that the subject's access rights to the object may change over time.

In addition, the entities used in the model represent some hierarchical structure described in the corresponding database. This relationship allows you to request access to an operation on an object from a higher entity.

Thus, such a three – dimensional array is schematically represented in Figure 2:

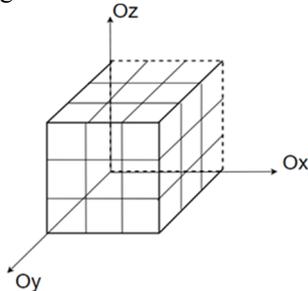


Fig. 2. SOT array.

In the course of a more detailed description of the model, consider the following questions:

1. The elements of the SOT array.
2. User right.
3. Using the subject database.
4. Principle of hierarchical relationships between subjects.
5. The models, with reference to the parameter t (time);
6. Security administrator rights.

The elements of the SOT array.

Note that in this case, under the SOT–array we will understand the element of three–dimensional space, which depends on three parameters: Subject (subject), Object (object) and Time (time).

Under the subject we will understand the essence of the computer system (person or process), the actions of which are governed by the rules of access control.

The object of access means such a unit of the automated information system resource, access to which is regulated by the rules of access control.

In addition to the two main elements of the known security models, we introduce the following element – time (t). Its application in this algorithm will be describe further.

User rights. In the described security model, subjects have the following rights over objects:

- a) w – write object (write);
- b) r – reading object (read);
- c) e – process activation (enable);
- d) i – request (inquiry).

In the three–dimensional security model, the "access matrix" looks like a parallelepiped whose axis Ox is represented by subjects, Oy –objects of the computer system, and the axis Oz –by time segments. However, just as in two–dimensional access matrix, the values written to the cells in the three–dimensional array define the types of safe access of the corresponding entity to the corresponding object.

By request, we will understand the situation, in which a subject, who does not have access to any subject, can send a request for temporary granting him the appropriate rights.

In other words, requesting one entity to another to perform operations on objects is as follows:

$$i (s_k ((o_1, x), (o_{2,x})) \rightarrow S_m,$$

where $n, k, n=1, 2, 3, \dots; x \in O: O = \{w, r, e\}$.

In this example, subject S_k principal sends a request to S_m to perform x operations on o_1, o_2, \dots, o_n objects. In this case, the set contains all possible operations on objects (write, read, enable).

The subject S_n then sends to S_k response to their request to perform operations on the objects:

$$S_m(f_{o_1, x}, f_{o_2, x}, \dots, f_{o_n, x}) \rightarrow S_k,$$

where $x \in O: O = \{w, r, e\}$,

$f = \{0, 1\}$ is the result of the query for the operation,

where 0 – failure, 1 – success.

Use of the database of subjects.

The axis of O_x of the three – dimensional massive is made of all subjects, which carry out working on computer system and over its objects in general.

Data about subjects, such as *logins* and *passwords* (if the subjects are meant to be real users of the computer system); *id* (if the subject is a process), as well as a list of parent and child in the hierarchy of subjects stored in the relevant database subjects on the server. It should be noted that to ensure security, the database stores directly the hash value of the login and password or id process.

This database can be represented as follows:

TABLE II. DATABASE

hash (s) OR hash (id)	z_1	z_2
hash (login ₁ +password ₁)	s_4, s_8	s_2, s_{13}
...
Hash (id ₁)	s_{15}, s_{23}	s_6, s_5

In this table the first column contain the hash values of usernames and passwords or id of the process, respectively. The second and third columns contain the related entities below and above in hierarchy.

During server operation, the SOT array is dynamically formed and adjusted, depending on the actions of the security administrator.

If for any reasons the specific subject loses an opportunity to make actions over objects from the SOT massif all mentions of this subject are removed. However, in the database all the entries including a hash value calculated based on the concatenation of the username and password are stored. It becomes for the purpose of prevention of a possible sort of the attacks initiated by this subject.

Principle of the hierarchical relations between subjects. Subjects of computer system are connected with each other by the hierarchical relations. Let us consider this principle on a concrete example, which is represented in the Figure 3.

The subject of s_1 with access to an object of o_1 wants to get access to an object of o_2 to which the subject of s_2 has an access. Apparently, from the drawing, these subjects are not connected by the hierarchy relation. Proceeding from it that the subject of s_1 could get access to a necessary object to it is required to send inquiry of i_1 to a higher subject with which it is connected by the hierarchical relations. In that case if the subject to which a request also was sent has no communication with o_2 object interesting to the subject s_1 the inquiry is sent further until is the subject having hierarchical communication with a necessary object of o_2 will find. The higher subjects can how to send inquiry further, or to reject it.

Possibilities of model with a binding to parameter t (time).

In this model of safety, the dependence of access on time can be realized the next ways.

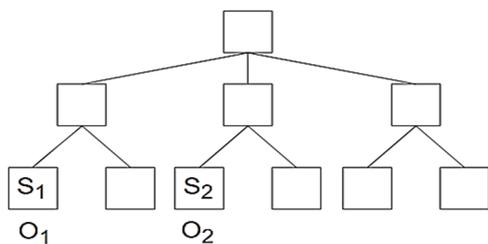


Fig. 3. Hierarchy of subjects.

The first of them is that the right of access to an object is defined by the actual time. In other words, the subject has access to an object to concrete temporary intervals. The current time is defined and stored on the server in encrypted form, and it in turn takes information from various interchangeable resources.

For example, the server time is 00:00. In the SOT–array it is specified that in the period from 00:00 to 10:00 the subject of s_3 has the right to record the O_3 object. Thus, in any period of time belonging to the specified it can perform operation w .

However, in the interval the subject not belonging to the described can perform other actions, if they are provided in the SOT–array. However, if this array has no information on the permissible operations of an entity on that object at this time, the right of access to it at the subject is absent completely.

The second method is that the subject's right of access to the object is not tied to server time, but is based on the principle of a timer. If it is specified that a subject have access to an object for a certain time t , then as soon as he has performed any action with it, the server activates a timer, after which access to the operations on the object can be changed or banned completely.

Security administrator rights.

The security administrator rights of the SOT model include the creation and adjustment of a three–dimensional array of access to objects by subjects. He can also modify the values of a database that contains data about the principals themselves, as well as their relationships to other model elements.

When changing the three–dimensional model, namely the time parameter (in the event that the subject is a real user of the system), the administrator must take into account such situations as:

- Weekends and holidays.
- The dismissal of an employee or the emergence of a new.
- Staffing changes.
- Promotion or demotion of a staff member, which provides excellent access to computer facilities.

Possible threats and weaknesses of the security model.

1. If a certain illegitimate person learns a valid login and password of a real user of the computer system, it may obtain the rights possessed by the subject.

2. To protect against various types of attacks, it is recommended to use software and hardware.

3. Dynamic change of SOT–array assumes the high requirements to the computing systems.

4. In the case of incorrect filling of the subject's database, looping may occur when sending a request to the subject standing higher in the hierarchical tree of relations. For example, the subjects in the inquiry will be to constantly refer to each other.

IV. CONCLUSION

In this article already, existing models of access were considered and their advantages and disadvantages based on which the new model of security of computer systems was offered are designated.

In addition, a graphical representation of the security model was provided for greater clarity. The database of subjects and their hierarchical relations was also described; a general description of the functioning of the algorithm of the security model was given.

Based on the above, possible security threats and shortcomings of the developed model were identified. The considered model is not final and can be subjected to various modifications.

REFERENCES

- [1] A. Chipiga, "Information Security of the Automated Systems". M.: Helios of ARV, 2010. – 336 p. (In Russian).
- [2] G. Weidman, "Penetration Testing: Hands – On Introduction to Hacking". M.: Helios of ARV, 2014. – 528 p.
- [3] M. Goodman, "Future Crimes: Everything Is Connected, Everyone Is Vulnerable, and What We Can Do About It". ARV. 2015. – 393 p.
- [4] P. Devyanin, "Model of Safety of Computer Systems. Management of Access and Informational Streams". M: GLT, 2013. – 338 p. (In Russian).
- [5] P. Devyanin, "Model of Safety of Computer Systems. Management of Access and Informational Streams: Manual for Higher Education Institutions". M.: GLT, 2012. – 320 p. (In Russian).
- [6] P. Devyanin, "Model of Safety of Computer Systems. Management of Access and Informational Streams: Manual for Higher Education Institutions". M.: The hot line – Telecom, 2016. – 342 p. (In Russian).
- [7] V.S. Alekseenko, , F.I. Akshentsev, O.B. Brown, "Model of Increase in Effectiveness and Safety of Production by Means of Perfecting of the Organization and Compensation". M.: Mountain Book, 2012. – 52 p.
- [8] T. Koppel, "Lights Out: A Cyberattack, Nation Unprepared, Surviving the Aftermath". 2015. – 279 p.
- [9] H. Deytel, P. Deytel, D. Chofnes, "Operating Systems". V.2. The Distributed Systems, Networks, Safety". M.: BINOMIAL, 2013. – 704 p.
- [10] N.A. Severtsev, "Systems Analysis and Model Operation of Safety". M.: The Higher School. 2006. – 462 p. (In Russian).
- [11] J. Stallings and N. Brown, "Computer Security: Principles and Practice", 3/e. Prentice Hall. 2014. – 820 p.
- [12] A. Cornflowers, A.Vasilkov, I. Vasilkov "Security and Management of Access in Information Systems: Manual". M.: Forum, Research Center INFRA. 2013. – 368 p. (In Russian).
- [13] A. Boyle and M. Panko, "Corporate Computer Security", 3/e. Prentice Hall. 2013. – 661 p.
- [14] J. Sammons, M. Cross, "The Basics of Cyber Safety: Computer and Mobile Device Safety Made Easy". Syngress. 1–st Edition. 2016. – 254 p.
- [15] A. Skavhaug, J. Guiochet, E. Schoitsch, F. Bitsch, etc. "Computer Safety, Reliability, and Security". SAFECOMP 2016 Workshops. 2016. – 400 p.
- [16] P. Devyanin, "Model of Safety of Computer Systems. Management of Access and Informational Streams. Educational". M.: GLT. 2012. – 305 p.

Interface and Software for the System of Automatic Seeding of Grain Crops

Maksim A. Litvinov
Federal Scientific Agro
Engineering Center VIM,
Moscow, Russia
litvvinov.max@yandex.ru

Maksim N. Moskovskiy
Federal Scientific Agro
Engineering Center VIM,
Moscow, Russia
maxmoskovsky74@yandex.ru

Ilya V. Pakhomov
Don State Technical University,
Rostov-on-Don, Russia
ilyavpakhomov@gmail.com

Igor G. Smirnov,
Federal Scientific Agro
Engineering Center
VIM, Moscow, Russia,
rashm-
smirnov@yandex.ru

Abstract— One of the sector of agriculture that needs automation in the first place is selection of seeds, because of stoop labor is used in most technical operations. We have designed software for an intelligent intellectual seeding system that implements the following functions: calculating the seeding rate for a plot length, measuring and monitoring the plot length error, calculating the frequency of rotation of a stepper engine using the averaged values of the rotation angles of two encoders, signaling the end of the cycle and the cassette cells, add / delete / change data on allotments in the built-in database. An algorithm has been developed to reduce errors during sowing long selection plots.

Keywords — automation agriculture, selection of seeds, software rate of sowing seeds, intelligent seeding system, algorithm

I. INTRODUCTION

In the world ranking, production of grain crops occupied leading positions [1], therefore, their production is a significant task. Most agriculture machines and equipment hasn't practically electronic automation systems and operating units are made in the form of standard kinematic gears and drives.

In the construction of most selection seeders, there are used both manual seed supply systems in sowing process and mechanical ones, which leads to an increase in losses of seed material, increased energy intensity and labor costs of the process.

In addition, the design and development of the applied selection machinery and equipment does not take into account such important characteristics as: soil-climatic features of the terrain, matching the speed of movement of the seeder with the speed of rotation of the sowing apparatus, taking into account the varying coefficient of skidding [2].

On the market of modern machinery of selection, we can find seeders with semi-automatic seeding systems, such as the Rowseed and Plotseed series from company «Wintersteiger» [3]. Its software contains a telemetry system, which show information about passed distance on the screen, a fixed database of seeding rates and manual settings for the mechanical drive of the sowing device, as well as an automatic control system for the sorting table and dispenser.

Also in India, scientists are developing of automatic systems for manual seeders [4]. This software works with following parameters: the presence of seeds in the bunker, obstacles on the way, and also tracking the end of the field by markers. This software is suitable for selection seeders of the 1st stage, but the its functional is insufficient for the 2nd, 3rd and 4th stages.

The benefits of this drive are control the correct distance between the seeds, to regulate the seeding rate. The drive is also portable and it can be used on small areas [4].

On this moment it isn't existed ESD electronic system with telemetric satellite system (Trimble system or Topcon system) and electro drive. ESD system is useful for application in large farms and in sowing of food grain due it is very expensive and its sowing device intended only for sowing roots tuber crops.

The aim of research is developing software for the implementation of an intelligent seeding system for selection seeders with the ability to control the optimal parameters of the technological process, reducing the distance error in sowing process at the long selection plots.

The novelty of the research is development of software that takes into account the error of distance during seeding, which occurs due to wheel skidding.

II. STRUCTURE OF THE PROCESS ALGORITHM

The technical objective of the software is to automation of the sowing seeds process and correcting the rotational speed of the sowing apparatus relative to the rotational frequency of the encoder disks installed on the driven wheels of the seeder, taking into account the difference of signals.

The software realizes the following functions: adding / deleting / changing information about plots with built-in database, signal about the end of the cycle or fully empty of cassette, calculating the plot length according to the encoder, calculating the rotation frequency of the stepper engine in accordance with the specified plot length, rotation frequency correction stepper engine based on the average value of the rotation angle of two encoders.

The software operates with the following information: the length of the plot, the number of cells in the cassette, the opening time of the dispenser, the angle of rotation of the stepping engine and the encoder.

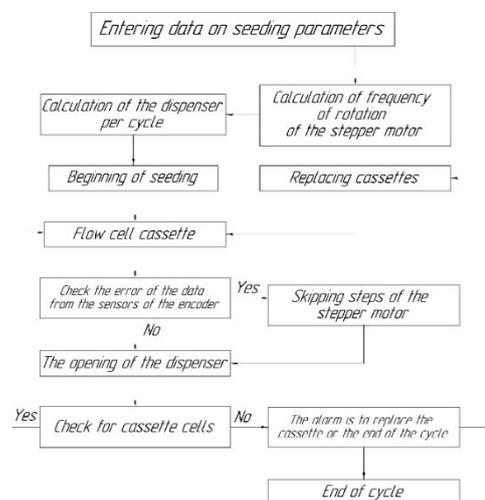


Fig. 1. Algorithm of software

The research is carried out at the expense of the State assignment № 10.9.05 (Reg. № AAAA-A18-118090390029-0)

The algorithm of software (Fig. 1) consists on several blocks: entering of information and calculation of seeding parameters, changing cassettes of the sorting table, and a block of executive device.

In the block of entering of information (Fig. 2), the operator fill fields of diameter of the seeders driven wheel, the length of the plot, the length of the dividing line, and the number of sections in the cassette.

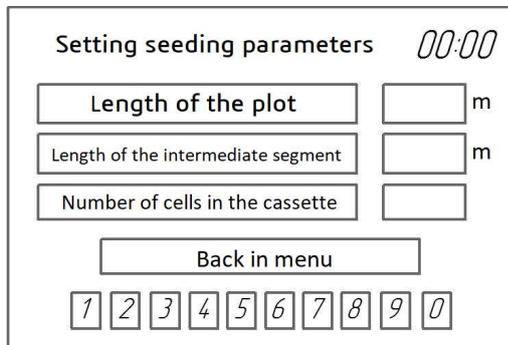


Fig. 2. Interface of block of entering information

After entering the initial information algorithm will signal the installation of sections of the cassette in a starting position. The operator will control the drive of the sorting table with help of interface (Fig. 3) until the switch is set on the table, which in turn is a section counter.

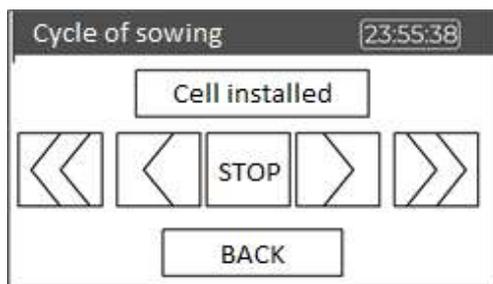


Fig. 3. Interface block of the installation of cassette section

The activation of the limit switch serves as a signal for the transition to the control executive devices. The graphical interface of the block shows (Fig. 4) the number of remaining sections of the cassette before changing, the distance for visual control of the rut length by the operator. In the field "current status" it shown the values: "sowing", "section change", "stop of sowing".

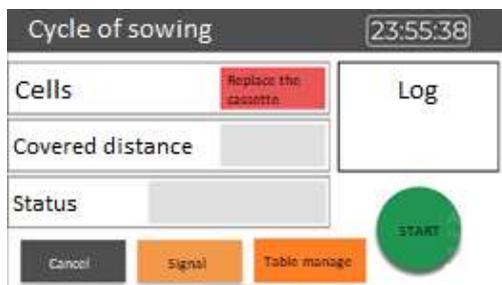


Fig. 4. Interface control of executive devices

III. SOFTWARE FOR SELECTION SEEDER

At this moment there are several world manufacturers of components for selection seeders: Amazon, Zürn, Maple, OEZ, Wintersteiger. It was taken components of the selection seeder Rowseed S, WINTERSTEIGER (distribution table with drive, sowing machine, grain metering unit) as a basis of the executive

devices, because this machine can be adapted to automatic seeding [6] by combining GSC and an intelligent seeding system.

The algorithm is based on the use an Arduino Mega 2560 microcontroller, which has the following characteristics: ATmega 2560 chip, 256 KB ROM, 16 MHz chip clock frequency, 70 digital I / Os (14 PWM), 16 analog inputs and 3 UART ports.

This controller corresponds to the stated characteristics and compared to the Iskra JS counterparts, the Raspberry Pi has a lower price and its own software development environment based on the C ++ language.

For showing the information we selected touch screen Nexion NX4827T043 - 4.3- display. The display has the following characteristics: resolution 480x272, TFT screen with integrated resistive touch screen with 4 wires, 16M Flash memory, power consumption 5V 250mA.

The smsd-4.2rs-485 driver was selected as the driver for the Nema 34 stepping engine; it provides all the necessary parameters specified in the specification to the motor.

It was used incremental encoder for forming signal. This signal allow identified relative offset from the previous position [7].The encoder was made in the form of a disk with rectangular teeth (Fig. 5), which was installed on the driven wheel, and optocouplers in the form of a laser and a photodiode.



Fig. 5. Encoder disk

In moving from two encoders a signal is generated (Fig. 6) and it is possible to determine the difference in the rotational speed of the driven wheels and also control the rotational speed of the stepping motor, thereby maintaining the required seeding rate.

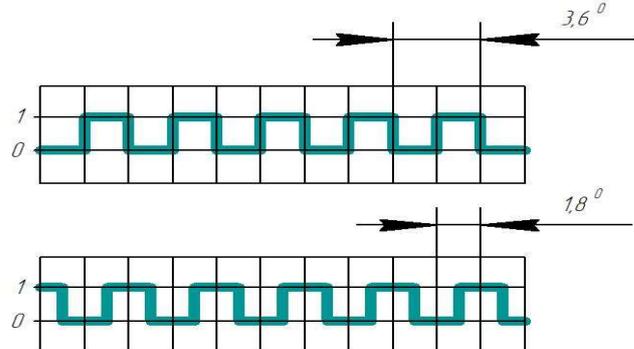


Fig. 6. Forms of signal from encoders

Designing code and graphical user interface is performed based on the Arduino IDE and Nextion Editor development environments [5].

IV. METHOD OF DESIGN

The following variables were introduced to calculate the seeding rate: D – diameter of the driven wheel in meters, d – diameter of sowing cone in millimeters (by default 122 mm), L – length of the plot, L_p – length of the intermediate section, P – number of sections, S – encoder step size, s – motor step size, N – number of encoder steps per revolution (constructively $N = 100$), n – number of steps of a stepper motor per revolution of the rotor (structurally $n = 200$).

The step length is calculated using the formula:

$$S = nD / N \quad (1)$$

The calculation of the number of steps of the encoder on the length of the plot L is made according to the formula:

$$N1 = L / S \quad (2)$$

When the parameter L is reached, the dispenser is given a signal to open, then a signal is sent to the distribution table to feed the cell. The counting of the number of sections P is carried out on the counter, when reaching the value 0, a message on the replacement of the cartridge is displayed on the screen and the cycle repeats.

Value of step size of stepper engine:

$$s = n * d / n \quad (3)$$

The motor step in relation to the encoder step is calculated from the ratio of the steps.

$$S1 = 2 * s * D / d \quad (4)$$

The calculation of the steps number of a stepper engine is made according to the formula:

$$n1 = L / S1 \quad (5)$$

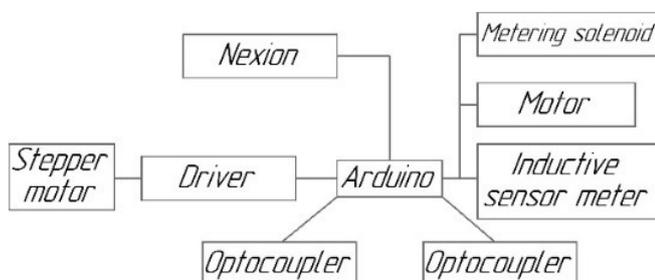


Fig. 7. Scheme of elements connection

The principle of operation of the system is that the optocouplers record dates from the reference disk (Fig. 5) in the pulses form. Based on the entered length of the plot, (2) and (5) we determine the required pulses number of the encoder steps and the stepper engine per plot. During driving, the speed of rotation of the stepping engine will slow down by the average value of the difference in the input signals from the sensors, when the difference in pulses from optocouplers will be 5%.

The stepper engine is rotated only during passing the plot with the dividing strips of the length L_p . The stepping engine is not rotated in the span of the seeder.

Before starting operation, the operator loads the cassettes into the separation table and places the first section by rotating the table motor forward or backward until the inductive sensor of the cell counter is triggered. From this position begins counting sections of the cassette.

After the tractor starts movement (arrival of impulses from the

encoder), the "start" button has to be pressed, the solenoid opens the dispenser for 2 seconds at the beginning of the plot, then the next cell is supplied and solenoid opens the dispenser for 2 seconds at the beginning of the plot. The cycle is repeated until the cells are ended, after a message is displayed on the screen of the operator to replace the cassette.

I. RESULTS AND DISCUSSIONS

Step engine Nema 34 have 200 steps for one turn, this fact will provide accuracy of step and accuracy is 1.80. Quantity of encoder step is 11 mm with relation encoder turns and step engine one to one. Driven wheel diameter is 700 mm and amount of encoder tooth's is 100 tooth's. Quantity of engine step is 0.22 mm accordingly standard length of plot is 100 mm in the second selection stage. This indicator is showed high quality of drive accuracy.

It was developed the prototype self-propelled planter based on the chassis VTZ-30 for field tests (Fig. 8). Testing of an automatic seeding system will be made in the period of sowing of winter cereals crops.



Fig. 8 Prototype self-propelled drills on the chassis VTZ-30

V. CONCLUSIONS

Graphical interface has been developed based on the simulation of the process and calculations of seeding parameters. It will allows the operator to record / change / monitor the required indicators and process data. Designed software will allows controlling the parameters of the seeding process by reducing the determining error in the distance of selection plots.

The peculiarity of the designed software is adaption to different types of encoders and stepper engines. Also it is universal and it can be used on different types of sowing systems.

REFERENCES

- [1] Clark, Andy, Cover crops / Sustainable Agriculture Research & Education (SARE) program, Handbook 9: 12-14, 2017, United States
- [2] Automatic system of intelligent seed rate control for selection seeders / 16th IEEE East-West Design & Test Symposium (EWDTS): 722-726, 14-17 September 2018, Kazan, Russia
- [3] <https://www.wintersteiger.com/ru/> PLANT CULTIVATION AND RESEARCH / Products / Assortment / Seeding Planters
- [4] Francis Perea. Arduino Essentials/ Packt Publishing, 2015, ISBN: 978-1784398569.
- [5] Ms. Trupti A. Shinde1, Engg. R.I.T., Sakhrle. Design and Development of Automatic Seed Sowing Machine / SSRG International Journal of Electronics and Communication Engineering - (ICRTESTM) - Special Issue – April 2017.
- [6] Frank M. Zoz, Robert D. Grisso. Traction and Tractor Performance / Agricultural Equipment Technology Conference, 9-11 February 2003, Louisville, Kentucky USA. ASAE Publication Number 913C0403
- [7] Mark Howard, General Manager. Incremental encoders, absolute encoders & pseudo-absolute encoders/ Technical whitepaper: 2017.

Cross–Platforming Web–Application of Electronic On–line Voting System on the Elections of Any Level

Evgeniy V. Palekha
Information System Cybersecurity Chair
Don State Technical University
Rostov–on–Don, Russia
palekha1994@mail.ru

Olga A. Safaryan
Information System Cybersecurity Chair
Don State Technical University
Rostov–on–Don, Russia
safari_2006@mail.ru

Larissa V. Cherkesova
Mathematics and Computer Sciences Chair
Don State Technical University
Rostov–on–Don, Russia
chia2002@inbox.ru
ORCID 0000–0002–9392–3140

Irina S. Trubchik
Mathematics and Computer Sciences Chair
Don State Technical University
Rostov–on–Don, Russia
trubchik@mail.ru

Vitaliy M. Porksheyan
Applied Mathematics Chair
Don State Technical University
Rostov–on–Don, Russia
spu–46@donstu.ru

Olga N. Manaenkova
Mathematics and Computer Sciences Chair
Don State Technical University
Rostov–on–Don, Russia
manaenkova_o@mail.ru

Sergey A. Morozov
Information Systems and Radioengineering
Don State Technical University
Rostov-on-Don, Russia
andrey@sssu.ru

Boris A. Akishin
Mathematics and Computer Sciences Chair
Don State Technical University
Rostov–on–Don, Russia
akiboralex@mail.ru

Abstract — *this paper presents the practical implementation of the electronic voting Web–system. Analysis of the developed algorithm was carried out, which was taken as basis in practical implementation of software product. Developed software product can be used for electronic voting at the elections of any level.*

Keywords: *electronic voting, elections of any level, software product, web – system, program application, El–Gamal encryption system, digital signature.*

I. INTRODUCTION

Currently, the conduct of voting by hand counting ballots is a rather laborious, resource-intensive and scrupulous process. The organization of elections at polling stations is time-consuming and may not always ensure the transparency and integrity of elections. In addition, some citizens, due to compelling circumstances, cannot arrive at the place of voting.

At this time, computers are used everywhere, so why not apply them to such an important event as the holding of elections? Moreover, usage of computer systems will not only make the voting process more convenient, but will also be able to protect it from falsification of results. Consider the system of electronic voting.

This system involves three parties: voters, moderators and administrator. None of the moderators is not able to replace the voice of a voter. It is possible to prevent change of result of vote by use of the strengthened version of cryptosystem El Gamal, and also by a ban of access of the administrator and moderators to this or that table of a database. Thus, it is possible to reduce the probability of vote rigging [1].

To spoof the results, the attackers will not only need to access the database, but also to hack into the cryptosystem used.

II. TASK DEFINITION

The application has been developed, which organizes the voting process. In order for the application to function correctly on all popular operating systems, it was decided to develop a web application, as it will provide maximum availability and eliminate the difficulties in the development for each individual operating system, whether Windows, Linux, MacOS, Android,

etc. The purpose of this application is to ensure the voting process and control over the voting process.

III. THEORETICAL BASE

Cryptosystem, which is used in this application, is based on an enhanced version of the scheme El–Gamal. The substantiation of the scheme complexity was given in the article by A.S. Mazurenko and N.S. Arkhangel'skaya [1].

Some parameters are generated according to this scheme, the rest parameters are generated by the administrator, except in specified cases. Inspectors are divided into teams t of people, considering that $n=kt$, where $n \in N$ is some natural number, $k \in Z_+$ – positive integer. Decryption requires the participation of t inspectors, where $2 \leq t \leq n$.

X secret key generation occurs. Next, divide x into secret shares, which will receive each of the results of the vote. The administrator then publishes the second part of the public key as $k_0=(p, g, y=g^x)$, which will be used by voters to encrypt their vote [2].

Some number of candidate ($v \in N$) candidates participate in the elections, voters can vote only for one of the candidates. The voter votes, his vote is encrypted, and the resulting cipher text is sent to the people checking the election results. After the voting, all the inspectors, who received all the cipher texts–voices, restore the secret key and decrypt the vectors – voices. Then all the vectors are summed up and the resulting vector is declared as the result of voting.

The complexity of hacking the constructed cryptosystem is equivalent to the complexity of solving the universally recognized difficult problem of Diffie–Hellman decision – making in the group G .

Although this cryptosystem is reliable, however, its practical implementation is extremely labor–intensive. Accordingly, only part of the scheme, namely the digital signature of El–Gamal, was used in the development of the web application. The third party participating in the scheme developed by A. Mazurenko and N. Arkhangel'skaya were inspectors, at practical implementation they were replaced

on moderators. El-Gamal's digital signature is applied by administrator to voter's unique identifier and voice.

IV. PRACTICAL IMPLEMENTATION

Web application is a client – server application in which a client accesses a web server through a browser and receives a response in the form of an HTML page [3].

Data is primarily stored on the server, and information is exchanged over the network. The developed application is based on the Model – View – Controller concept [4].

The General scheme of MVC is shown in the Figure 1. This is scheme for dividing the application data, user interface and logic of his work on three components:

- Model provides data and responds to controller commands by changing its state;
- View displays the resulting data from the model to the user; react to changing the model;
- Controller is a link between model and view, notifies model of user's response.

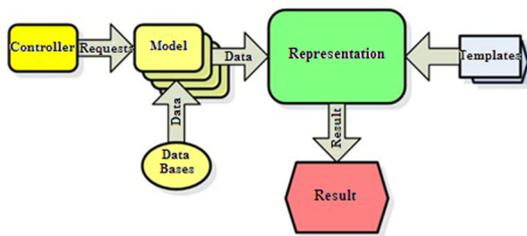


Fig.1. General Scheme of the MVC Concept.

This application uses one of the concept modifications – hierarchical, namely HMVC [4]. It is typically used with object – oriented language tools. Its essence lies in the fact that for each of the concepts three main classes are described: a common model class, a common controller class and a common presentation class [5].

Next, the development describes the new models that inherit the properties and methods of the General class of the model. Similar actions apply to new model and view classes. Consider the root directory of the application that contains the project, it is shown in the Figure 2.

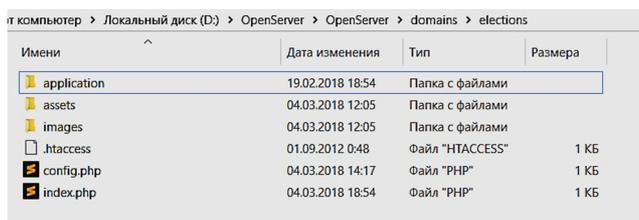


Fig. 2. Project root directory.

The MVC concept implies one entry point – the index.php. This file is the backbone of the project, through this script will pass all requests to the web application and all the logic of the project. In order to implement this approach, you must configure the server. It is assumed that the site runs on the apache server, this requires a file.ht access, which contains URL routing rules.

Routing will also allow you to create human – friendly URLs, so that the address of the pages will look like: project.ru/view/action Oh. This example calls the action method of the view controller class.

Config.php is a project configuration file that contains data, such as the full project path, as well as database connection constants (user, password, host, database name). The assets folder contains style files and js scripts, and the images folder obviously contains images for the project.

The core of the project is the application folder – it contains the whole MVC concept structure, the contents of the application folder are shown in the Figure 3.

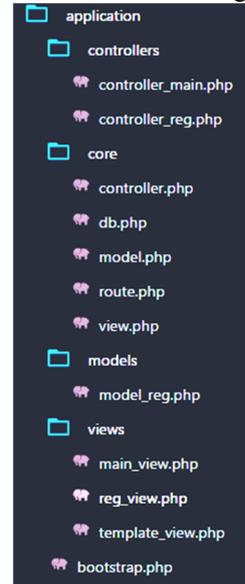


Fig.3. Application directory content

This folder contains the bootstrap.php. It collects all the files from the project core. All files from the folder core is connected to the bootstrap file.php and called the router, which will guide the application. The router class is described in the route.php and shown in the Figure 4.

The task of the router is to select the desired controller from the address bar, call the specified method of this controller and provide to user the output result using the specified view [6].

```

// контроллер и действие по умолчанию
$controller_name = 'Main';
$action_name = 'index';
$routes = explode('/', $_SERVER['REQUEST_URI']);

// получаем имя контроллера
if(!empty($routes[1]))
{
    $controller_name = $routes[1];
}
if(!empty($routes[2]))
{
    $action_name = $routes[2];
}

// добавляем префиксы
$model_name = 'Model_'.$controller_name;
$controller_name = 'Controller_'.$controller_name;
$action_name = 'action_'.$action_name;

// подключаем файл с классом модели (файла модели может и не быть)
$model_file = strtolower($model_name).'.php';
$model_path = 'application/models/'.$model_file;
if(file_exists($model_path)){
    include 'application/models/'.$model_file;
}

$controller_file = strtolower($controller_name).'.php';
$controller_path = 'application/controllers/'.$controller_file;
if(file_exists($controller_path))
{
    include 'application/controllers/'.$controller_file;
}
else
{
    exit('Файл контроллера не найден');
}

// создаем контроллер
$controller = new $controller_name;
$action = $action_name;
if(method_exists($controller, $action))
{
    // вызываем действия контроллера
    $controller->$action();
}
  
```

Fig. 4. Route class code.

To simplify code development and readability, controller, model, and view names have the same names, except for prefixes. It is also taken into account that the address can be deliberately changed and the desired controller does not exist, it uses the default value, or there is no specified method, then an error message is displayed. It is worth noting that the model file may not be, and is required only the presence of the controller.

The Controller General class contains a constructor that declares the view class. It also describes the action index () abstract method [6], which is required for each individual controller to perform its tasks during inheritance.

The description of the General view class contains only the generate method (\$content_view, \$template_view, \$data = null). This method connects the \$template_view page template, the content of which is filled with the \$content_viewpage, and the \$data parameter is an array that contains the data retrieved from the model [6]. It is worth noting that the default page template is template_view.php, the code of which is described in Figure 5.

```
<!DOCTYPE HTML>
<html>
  <head>
    <title>Выборы лучшего кандидата</title>
    <meta charset="utf-8" />
    <meta name="viewport" content="width=device-width, initial-scale=1" />
    <!--[if lte IE 8]><script src="<?-SITE_PATH>assets/js/ie/html5shiv.js"</script><![endif-->
    <link rel="stylesheet" href="<?-SITE_PATH>assets/css/main.css" />
    <!--[if lte IE 8]><link rel="stylesheet" href="<?-SITE_PATH>assets/css/ie8.css" /><![endif-->
  </head>
  <body>

    <?php include 'application/views/'.$content_view; ?>

    <!-- Scripts -->
    <script src="<?-SITE_PATH>assets/js/jquery.min.js"</script>
    <script src="<?-SITE_PATH>assets/js/jquery.scrollTo.min.js"</script>
    <script src="<?-SITE_PATH>assets/js/jquery.poptrox.min.js"</script>
    <script src="<?-SITE_PATH>assets/js/skel.min.js"</script>
    <script src="<?-SITE_PATH>assets/js/util.js"</script>
    <!--[if lte IE 8]><script src="<?-SITE_PATH>assets/js/ie/respond.min.js"</script><![endif-->
    <script src="<?-SITE_PATH>assets/js/main.js"</script>

  </body>
</html>
```

Fig. 5. Code of the default template.

As you can see from Figure 5, the page template contains only the General structure of the HTML document, as well as the connection of the style file and js scripts. Also connects the contents of the required page from \$content_view. A generic model class consists only of the get_data method, which is an abstract method. It also describes a single General method, which can be attributed to the model, so it is in it that the database is accessed. The db.php contains a description of this class, DB class is abstract, but it contains a lot of methods that will be useful to all classes that inherit these methods.

Using these methods, you can perform various database manipulations without using SQL in its pure form [7]. It is enough to describe a new class, the name of which will tell the General DB class which table from the database to work with. You also need to specify what conditions are needed to extract data from the database. For example, on which of the columns to conduct a search, or you can specify a limit on the issuance results, etc.

Therefore, on the main page of the application contains a list of candidates that voters can choose. This page is the default page whose controller is described in the

controller_mainfile.php (this controller does not have its own model works independently). This file describes the Controller_main class, whose code is shown in Figure 6.

```
<?php
class Controller_main extends Controller
{
  function action_index()
  {
    $model = new DB_Candidates(); // создаем объект модели
    $usersInfo = $model->getAllRows(); // получаем все строки
    $this->view->generate('main_view.php','template_view.php', $usersInfo);
  }
}
```

Fig. 6. Class description Controller_main.

Here selects all the candidates in the voting from the database and calls the main_viewview.php. This file displays information about candidates and their brief description.

To participate in the voting, the voter must register. To do this, click on the link "Register". During the transition, there is a change of controller on Controller_Reg, which causes the registration form. When registering, you must specify the full name, date of birth, passport details, date of issue and TIN.

After filling in the data and sending the form, there is a void check, as well as checking the correctness of the entered dates, as well as series and passport number.

The passport series consists of four digits, the first of which contains information about the region in which the passport was issued. The region code is checked by the list, for example, Moscow has the code 45, and Rostov region corresponds to the code 60. This check is necessary in order to avoid a possible attempt to falsify the voting results by introducing a non-existent region code. The second 2 digits of the passport series indicate the year of issue of the document. This value can not be more than 18 or less than 97, as it was October 1, 1997 passports of the USSR began to replace on passport of the Russian Federation.

The passport number is 6 digits, and paired with a series is a unique identifier of the voter, matches in the database with the same series and the passport number is impossible, so voters are tested for matches to ensure that the voter could not vote more than once. The TIN must also be filled in according to the rules and consist of 12 digits.

Also, when registering, you must choose, in fact, your candidate. It uses a separate table in the database, which contains two columns: voter id, candidate id. They point to each other's conformity. It is worth noting that none of the moderators or even the administrator will not be able to change the values of this table, because they do not have privileges to modify the data in this table, only to add data [8].

For additional data protection, a special table is used, which also contains two columns: the voter id and the result of the El Gamal digital signature [9].

It is the administrator who signs the concatenation from the ID of the candidate and the series, and the passport number of the voter with his private key. This table has several purposes. The signature is used to verify the results of the vote, whether the voter actually chose this particular candidate. This check is necessary only when attempting to compromise the electronic voting system [10].

To do this, an attacker must have access to the root super user and replace the votes in the table. After the registration is complete, the voter goes to the voting results page. The results of the voting are presented in the form of a pie chart, which can be seen in Figure 7.

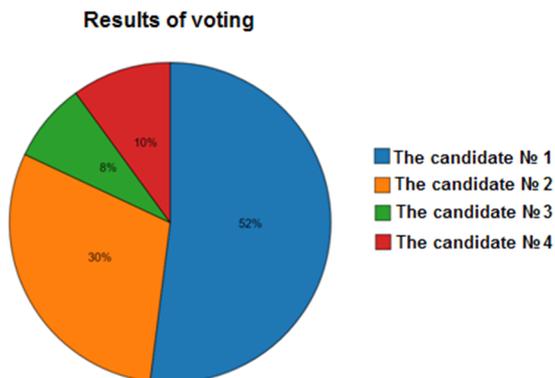


Fig. 7. Election Results.

Before counting votes, the system of voting checks whether the signatures match, and only after that the results of voting are displayed.

It should be noted that there may be cases when the voter accidentally entered his passport data incorrectly, thereby prohibiting the voter from voting, whose data he entered. This requires moderators who can change the data in the voter table. After changing the data on the series and passport number, the administrator again signs the concatenation of the id of the selected candidate and the series, and the passport number of the voter [11 – 13].

V. CONCLUSION

Thus, an enhanced version of the El Gamal cryptosystem was analyzed for the practical implementation of the electronic voting system. As a result, a web-based application for electronic voting has been developed, which ensures the simplicity, honesty and transparency of voting. The developed application functions correctly on all popular operating systems. In addition, this application contains protection against data spoofing, thus significantly reduces the likelihood that the system will be compromised [14], [15].

REFERENCES

- [1] A. Mazurenko, N. Arhangel'skaja, L. Cherkesova, O. Safaryan "Computational Complexity of Coding and Information Security System Based on Threshold Secret Sharing Scheme Used for Electronic Voting Systems". – Rostov-on-Don: DSTU, 2017. (In Russian).
- [2] R. Barbulescu, P. Gaudry, A. Joux, E. Thom'se, "A heuristic quasi-polynomial algorithm for discrete logarithm in finite fields of small characteristic". In: P.Q. Nguyen, E. Oswald, (eds.) EUROCRYPT 2014. LNCS, Springer, Heidelberg (May 2014). Vol. 8441, Pp. 1–16.
- [3] Free encyclopedia [Electronic resource]. – Access mode: <https://ru.wikipedia.org/wiki> (circulation date 25.02.2018).
- [4] S. Rogachev, "A Generalized Model-View-Controller". – Moscow: K-Press, 2016. – Pp. 37–66. (In Russian).
- [5] D. Kolisnichenko, "PHP and MySQL. Development of Web applications". – St. Petersburg: BHV-Petersburg, 2014. – 560 p. (In Russian).
- [6] S. Prettyman, "Learn PHP 7: Object Oriented Modular Programming using HTML5, CSS3, JavaScript, XML, JSON, and MySQL". Apress; 2015. – 294 p.
- [7] W. Stallings "Cryptography and Network Security", 7Th Edition". Pearson India; VII-th Edition, 2016.
- [8] J.-P. Aumasson, "Serious Cryptography: Practical Introduction to Modern Encryption", 2017.–312 p.
- [9] V. Stone, S. James, "Information Theory: A Tutorial Introduction". Sebel Press, 2015. – 260 p.
- [10] Ye. Lindell, "Introduction to Modern Cryptography. Chapman and Hall / CRC", II edition (November 6, 2014). 603 p.
- [11] S.A. Zheltov, "Effective Computing in the CUDA Architecture in Information Security Applications". PhD dis. / Zheltov S.A. – M: IINTB RSUH, 2014 – 145 p. (In Russian).
- [12] P. Kocher, "Timing attacks on implementations of Diffie- Hellman, RSA, DSS, and other systems", Advances in Cryptology Crypto96, Springer-Verlag, LNCS 1109, 1996, Pp.104–113.
- [13] J. Sammons, M. Cross, "The Basics of Cyber Safety". Computer and Mobile Device Safety Made Easy". Syngress. I-st Edition. 2016. – 254 p.
- [14] A.F. Chipiga, "Information Security of the Automated Systems". / A.F. Chipiga. – M.: Helios of ARV, 2010. – 336 p.
- [15] T. Koppel "Lights Out: A Cyberattack, Nation Unprepared, Surviving the Aftermath". 2015. – 279 p.

Theoretical Bases of the Course Motion Two Axles Agriculture Transports Vehicle According Wheels Slipping

Maksim A. Litvinov
Federal Scientific Agro
Engineering Center VIM,
Moscow, Russia
litvvinov.max@yandex.ru

Maksim N. Moskovskiy
Federal Scientific Agro
Engineering Center VIM,
Moscow, Russia
maxmoskovsky74@yandex.ru

Ilya V. Pakhomov
Don State Technical University,
Rostov-on-Don, Russia
ilyavpakhomov@gmail.com

Anatoly A. Gulyaev
Federal Scientific
Agro Engineering
Center VIM,
Moscow, Russia
Tomass1086@mail.ru

Abstract—Problem of linear motion of agricultural transporting vehicle (ATV) without satellite systems is complex difficult task. The consequences of non-linear movement associated with irregularity fertilizing. Non-compliance between the rows and irregularity plowing leads will leads to the unnecessary costs for operating materials. Also It will cause significant economics losses to agricultural enterprises. In this case, it was presented the theoretical foundations of simulation modeling of the course motion of the ATV and it was taking into account wheel slipping for analyzing the effects of non-linear motion. It was presented some results of checking their adequacy from field tests.

Keywords — automation agriculture, agricultural transporting vehicle, simulation modeling of the course motion, system of course stability, agricultural robots

I. INTRODUCTION

Important problem of the automation agricultural sector is auto piloting of agriculture machinery, but it is necessary to ensuring the linearity of movement due the loss communication with the satellites. Accidents of the automobile transport is one of the most actual socio-economic problems of most countries.

In the atmosphere of the high growth automation level of agricultural transports, the linearity of motion will become one of the most serious socio-economic problems. Successful solution of this problems depended on not only the reduction in labor costs, but also the development of the country's economy.

It should be noted that over the past seven years, the agricultural machinery park has noticeably lost in its quantity [1].

Particularly, the quantity of tractors had decreased from 174 287 units (2012) to 125 134 (2018), and the combine harvesters had decreased from 41 581 units to 3 313 units. Also in this list of agriculture machinery we saw machines with age of exploitation which more than 10 years. In these condition it is necessary more actual quality analysis of control of this machines.

In such situation, it is necessary specified analytical equations for describing motion of agricultural transporting vehicle (ATV) for researching and expertise special technological operations.

These specifies must consider «non-linear» of the processes, which connects with steering-wheel play and stiffness of steering linkage, transverse pitch of frame and also such the phenomenon's as «slipping» and «skidding». At the same time, the processes of curvilinear motion are insufficiently studied in operation manipulation with variable motion speed.

It is presented researches in the paper [2]. Authors, Danwei Wang и Feng Qi, shown modelling of the course motion of all wheel drive transporting vehicle.

The research is carried out at the expense of the State assignment № 10.9.05 (Reg. № AAAA-A18-118090390029-0)

Authors of paper developed model of course stability on the kinematic bicycle scheme with two drive wheels. This method of evaluation of course stability is useful for modelling of vehicle with four driven wheels, but using of this model will be incorrect for vehicles with two driven wheels as it wont take into account friction forces uncontrolled wheels in turn and drift rear axle.

Kinematic of control device and transporting vehicles dynamic is more exactly shown in Compendium for Course MMF062 [3].

Author accounted all forces acting on transporting vehicle, as example it may be automobile. Therefore, the basis for the development of a model of course stability of an agricultural vehicle will take the methodology presented in this book.

Aim of the research is development theoretical basis for calculating course movement of the two-axle agricultural transporting vehicle with regard to wheels slipping.

The novelty of the research is development new system of course stability without relation to geolocation. The results of research are recommended for creating agricultural robots.

II. THEORETICAL BACKGROUND OF MODEL DESIGN.

The properties of the mechanical system (ATV) don't depend on the choice of reference system.

That's why, it is proposed the most intrinsic choice of coordinate systems It isn't initially related to the specific properties of the system: the position of the mass center and the kinematics of the undercarriage:

–axis XO is the horizontal longitudinal axis. This axis connected to the ATVs frame of the located in the road plane;

–axis YO is a horizontal transverse axis. This axis connected to the ATVs frame located in the road plane and passing through the front axle of the frame;

–the axis ZO is the vertical axis. This axis connected with the ATVs frame, located in the vertical plane passing through the front axis of the ATV.

In this variant, we have possibility of design equations for limited values of the angular and linear rates of the ATV. We have the possibility of researching the course motion of the ATV with arbitrary values of its angular and linear displacements.

In the basis $\{Y_A, X_A\}$ for generalized coordinates, differential equations of motion can be written using any analytical method (for example, the Lagrange equations of the 2nd order); (Fig.1):

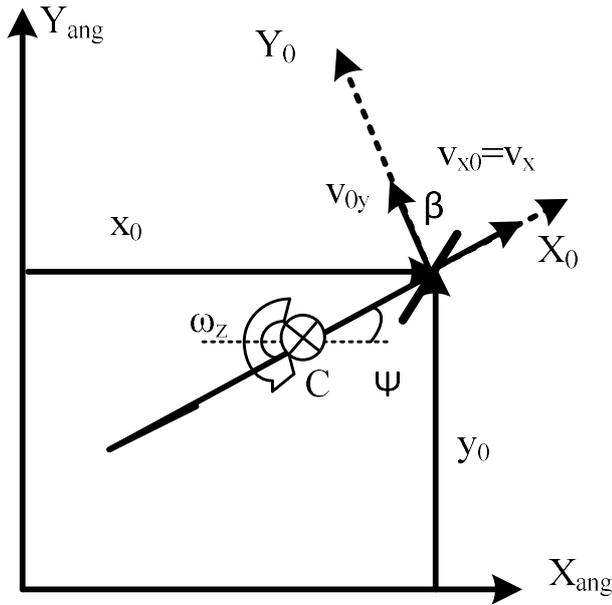


Fig. 1. Scheme and differential equations of the ATV course motion

$$m \cdot \ddot{y}_o - m \cdot L_1 \cdot \ddot{\psi} = R_1 + R_2; \quad (1)$$

$$-m \cdot L_1 \cdot \ddot{y}_o + I_{z1} \cdot \ddot{\psi} = -R_2 \cdot (L_1 + L_2); \quad (2)$$

$$\dot{y}_o = V_x \cdot \psi + V_{cy} + L_1 \cdot \dot{\psi}; V_{oy} = V_{cy} + L_1 \cdot \dot{\psi}; \quad (3)$$

$$\ddot{y}_o = V_x \cdot \dot{\psi} + \dot{V}_{cy} + L_1 \cdot \ddot{\psi}; \ddot{y}_o = V_x \cdot \dot{\psi} + \dot{V}_{oy}; \quad (4)$$

$$\delta_1 = \frac{V_{oy}}{V} - \beta; \quad R_1 = -k_1 \cdot \delta_1; \quad (5)$$

$$\delta_2 = \frac{V_{oy} - (L_1 + L_2) \cdot \omega_z}{V}; \quad R_2 = -k_2 \cdot \delta_2; \quad (6)$$

After completing the transformations, we will obtain the following algorithm for composing the equations of ATV motion with one cyclic coordinate ψ (heading angle), and according to this it was selected the basis coordinates of the system state (by the number of degrees of freedom) [2,6] and pseudo-speed vector - basic vector.

$$\bar{\mu} = \begin{bmatrix} y_o \\ \psi \\ \phi \\ \beta_1 \\ \beta_2 \\ \theta \\ x_o \end{bmatrix} = \begin{bmatrix} \text{transverse motion of the directional point} \\ \text{course_angle} \\ \text{transverse_pitch} \\ \text{angle of rotation of steering wheels -1} \\ \text{angle of rotation of steering wheels -2} \\ \text{rotation_of_driving_wheel} \\ \text{longitudinal motion of the directional point} \end{bmatrix} \quad (7)$$

$$\bar{\eta} = \begin{bmatrix} V_{oy} \\ \omega_z \\ \omega_\phi \\ \omega_{\beta 1} \\ \omega_{\beta 2} \\ \omega_\theta \\ V_{ox} \end{bmatrix} = \begin{bmatrix} \text{transverse speed of the directional point} \\ \text{rotational_speed_yaws} \\ \text{rotational speed of_transverse pitch} \\ \text{rotational speed of_steering_wheel -1} \\ \text{rotational speed of_steering_wheel -2} \\ \text{rotational speed of rotation of driving wheel} \\ \text{ground speed of ATV} \end{bmatrix} \quad (8)$$

Kinematic and power parameters of the ATV are shown in Fig. 2 at the course motion.

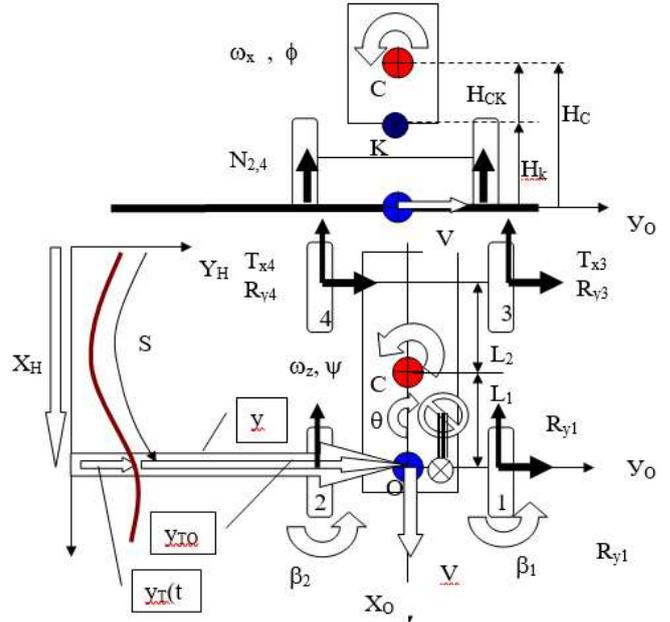


Fig. 2. Kinematic and power parameters of the ATV

III. STRUCTURE OF GLOBAL MATHEMATICAL MODEL OF ATV.

Vector of system condition was accepted in such way:

$$\bar{q} = \begin{bmatrix} \eta \\ \mu \end{bmatrix}. \quad (9)$$

In accordance with the aim, there are no restrictions on all degrees of freedom in the course motion, this means that ATV position can be absolutely unrestricted on the plane.

The transition from «pseudo-speeds» to absolute coordinates was made on the basis of an auxiliary system of differential equations:

$$\begin{bmatrix} \dot{x}_o \\ \dot{y}_o \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} \cos(\psi) & -\sin(\psi) & 0 \\ \sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} V_{ox} \\ V_{oy} \\ \omega_z \end{bmatrix} \quad (10)$$

The resulting system was reduced to normalized form for the system of differential equations of the first order for design evaluating algorithm based on standard programs for solving systems of differential equations:

$$\dot{\bar{q}} = \left\{ \begin{array}{l} A^{-1} \cdot \bar{Q} \\ [D_\psi(v, \psi)] \cdot \bar{q} \end{array} \right\} \quad (11)$$

where A is inertial matrix;

A-1 is inverse matrix;

$[D_\psi(v, \psi)]$ is transition matrix to generalized coordinates from generalized speeds;

\bar{Q} is vector of generalized forces, which consists of the following components:

$$\bar{Q} = \bar{Q}_{\lambda\phi} + \bar{Q}_{\beta\psi} + \bar{Q}_\phi + \bar{Q}_\psi + \bar{Q}_{\beta k} + \bar{Q}_{T_k} + \bar{Q}_{R_y} \quad (12)$$

The content of the individual components of the equation (14) for their evaluating, are given in the papers of the author as well as the basic designations [3,8].

Thus, the global structure of the ATV mathematical model was formed in the following form:

$$\{E_I, D_{\psi}(v, \psi), X_{pk}, X_N(\beta_1, \beta_2, \theta), X_{pk} X_{\phi}, G_{\delta}(v), C_R, M_{Ry}(\delta, N_{dim}), M_K(\delta, N), Ry_K(\delta, N), R_y, P_{RM}, N_{dim}, M_{cp}(\beta_1, \beta_2, \theta), M_{pk}(t), F_{py} Tx(t), M_{\beta}(\beta, N_{dim})\}. \quad (13)$$

In the final form, the mathematical model of the ATV course motion was described on the basis of the matrix expressions [4,9], shown in Fig. 3.

Accounting of ATV lateral slipping.

Accounting for slipping relative to the field carried out as follows [5,7]:

- action of the reactive moment from the field surface on the steered wheels at the point of contact (taking into account partial and full slipping) was shown such:

$$M_K(\delta, N) = \frac{N_{dim} \cdot k_m \cdot e^{-C_{\delta} \cdot \delta^2}}{C_{\delta} \cdot N_{z10} \cdot \sigma} \arctg(C_{\delta} \cdot \delta) \quad (14)$$

- transverse reaction from the field surface to the wheels (including partial and full slipping) was shown such:

$$Ry_K(\delta, N) = \frac{N_{dim} \cdot k_{\delta}}{C_{\delta} \cdot N_{z10} \cdot \sigma} \arctg(C_{\delta} \cdot \delta) \quad (15)$$

- kinematic coefficients of the wheels drift structure (the remaining zero elements of the matrix are not noted) were shown such:

$$G_{\delta}(v) = \begin{matrix} \frac{1}{v} & \frac{L_1 \cdot (1-SK)}{v} & \frac{-H_{\delta}}{v} & \dots & -(1-SK) & \lambda_1 & -(1-BK) & 0 & \dots & -BK \\ \frac{1}{v} & \frac{L_2 \cdot (1-SK)}{v} & \frac{-H_{\delta}}{v} & \dots & -(1-SK) & \lambda_1 & 0 & -(1-BK) & \dots & -BK \\ \frac{1}{v} & \frac{-L_1 - L_2 \cdot SK}{v} & \frac{-H_{\delta}}{v} & \dots & -(1-SK) & \lambda_2 & 0 & 0 & \dots & 0 \\ \frac{1}{v} & \frac{-L_1 - L_2 \cdot SK}{v} & \frac{-H_{\delta}}{v} & \dots & -(1-SK) & \lambda_2 & 0 & 0 & \dots & 0 \end{matrix} \quad (16)$$

A fragment of the maneuver “entrance into a turn - a turn from an obstacle” animation was shown in Fig. 4.

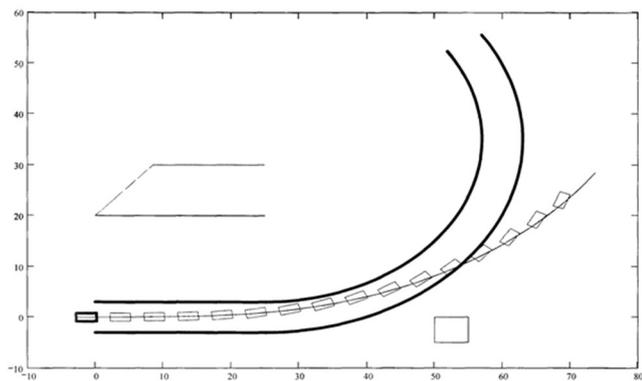


Fig. 3. Animation of the maneuver “entrance into a turn - a turn from an obstacle” at a speed of 12 km / h and the slipping.

IV. RESULTS AND DISCUSSIONS

The adequacy of mathematical models was tested on the basis analysis in the program MATLAB. It was illustrated in Table. 1 and Fig. 5.

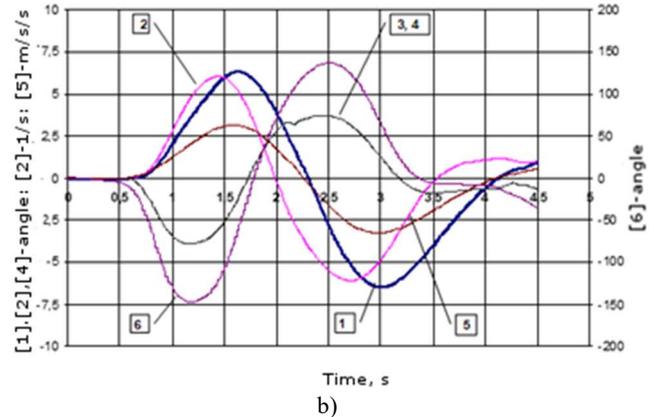
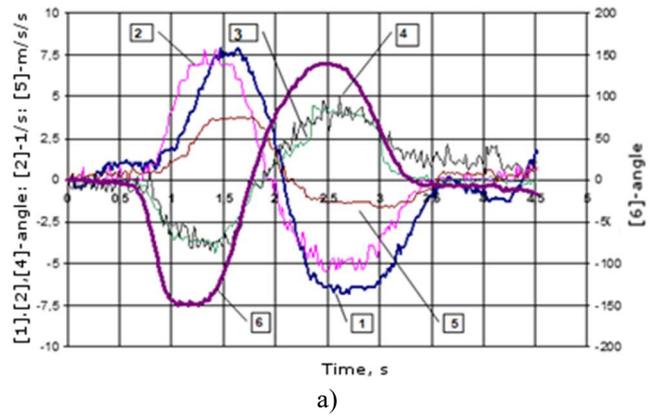


Fig. 5. Dependencies of parameters change: 1 - lateral pitch of frame, 2 - angular speed of ATV, 3 and 4 - right and left angles of rotation, 5 - lateral acceleration, 6 - angle of rotation of driving wheel: a) - field tests, b) - simulation modeling of the “reset” maneuver (length 20 m, width 3.5 m, speed ATV - 12 km / hour)

TABLE I. - RESULTS OF CHECKING ADEQUACY OF THE MODEL OF THE COURSE MOTION

Parameter maneuver	Model adequacy check	Angle of cross roll, ang.	Angular speed of ATV, rad/s	Right angles of rotation, ang.	Left angles of rotation, ang.	Angle of rotation of driving wheel, ang.	Lateral acceleration, m/s ²
Harshness of driving wheel	1500 N/rad.	Testing at the landfill	7.36	0.33	4.44	4.57	145.3
Clearance of driving wheel	10 ang.						
Lateral movement in maneuver	4.02 m.	simulation modeling	6.40	0.30	3.83	3.83	141.8
Speed ATV	12 km/h	Discrepancy, %	12.9	8.6	13.7	16.0	2.4

V. CONCLUSIONS.

Designed mathematical apparatus can be use for creating software for agriculture self-propelled robots. This apparatus will be accept complicated physical process of dynamic of curvilinear motion ATV, full or partial wheel slip, «slipping» and «skidding». It is especially important for maintaining the linearity of movement in the field.

Mathematical model also could be use as auxiliary program for adjustments steering in such systems as Trimble и LD-Agro UniDrive. These systems are focused mainly on satellite navigation.

REFERENCE

- [1] <https://www.fedstat.ru/indicator/33410>
- [2] Danwei Wang Feng Qi. Trajectory Planning for a Four-Wheel-Steering Vehicle/ School of Electrical and Electronic/ Engineering Nanyang Technological University/May 21-26, 2001/ Singapore.
- [3] Bengt Jacobson./ Vehicle Dynamics/ Göteborg, 2015
- [4] Zhileikin M.M./ The theoretical basis for improving sustainability performance, and manageability-wheeled vehicles on the basis of fuzzy logic techniques/ Bauman Press: 2016, p. 238, Moscow.
- [5] O. Nazarko, V.Boldovsky / Evaluation of the stability of the vehicle against skidding in a traction mode using computer/ Road transport: vol. 31, 2012, p. 26-27, Moscow
- [6] Shadrin S., Ivanov A./ Algorithm of autonomous vehicle steering system control law estimation while the desired trajectory driving/ ARPN Journal of Engineering and Applied Sciences: 2016, vol. 15 p 9312-6.
- [7] Buznikov S.E., Elkin D.S., Shabanov N.S. and Strukov V.O./ Task of safe automatic braking of the vehicle/ Trudy NAMI: 2016, p. 44-52, Moscow.
- [8] Jeong E., Oh C./ Evaluating the effectiveness of active vehicle safety systems/ Accident Analysis & Prevention: 2016, p. 85-96.
- [9] O. A. Sushchenko, A. A. Tunik/ Robust Stabilization of UAV Observation Equipment/ IEEE 2nd Int. Conf. Actual Problems of Unmanned Air Vehicles Developments October: 15–17, 2013, pp. 176–180.

Modification and Optimization of Solovey–Strassen’s Fast Exponentiation Probabilistic Test Binary Algorithm

Nikita Ye. Myzdrikov
Information System Cybersecurity Chair
Don State Technical University
Rostov–on–Don, Russia
nmyzdrikov@mail.ru

Olga A. Safaryan
Information System Cybersecurity Chair
Don State Technical University
Rostov–on–Don, Russia
safari_2006@mail.ru

Larissa V. Cherkesova
Mathematics and Computer Sciences Chair
Don State Technical University
Rostov–on–Don, Russia
chia2002@inbox.ru

Ivan Ye. Semeonov
Applied Mathematics Chair
Don State Technical University
Rostov–on–Don, Russia
ivan.sk6@gmail.com

Irina V. Reshetnikova
Communication Systems in
the Railway Transportation Chair
Rostov State University of
the Means of Communication
Rostov–on–Don, Russia
irina-rostov@mail.ru

Vasiliy I. Yukhnov
Info communicative technologies
and communication systems Chair
Moscow Technical University of
Communication and Computer Science
Rostov–on–Don, Russia
yuchnov@mail.ru

Andrey G. Lobodenko
Information Systems and Radioengineering
Don State Technical University
Rostov–on–Don, Russia
andrey@sssu.ru

Vitaliy M. Porksheyev
Applied Mathematics Chair
Don State Technical University
Rostov–on–Don, Russia
spu-46@donstu.ru

Abstract— This article will consider the probability test of Solovey–Strassen, to determine the simplicity of the number and its possible modifications. This test allows for the shortest possible time to determine whether the number is prime or not. C# programming language was used to implement the algorithm in practice.

Keywords: probability, test, algorithm, Solovey–Strassen, Jacobi symbol, binary exponentiation, Fermat, Fermat's little theorem.

I. INTRODUCTION

For many public – key cryptosystems critical for their full-time job, is a steady stream of prime numbers. For example, to use the Diffie – Hellman protocol and similar algorithms, it is necessary to generate a prime number p , which is the size of the residue ring field F_p . Another good example is the El–Gamal cryptosystem with a public key, in this algorithm prime numbers are needed to create a public key [1].

It should be noted that for the quality and smooth operation of many crypto–systems and protocols need a high–speed generation of the simplest numbers [2].

On the other hand, there may be a question of whether they should be generated at all, if they can be counted at once and stored anywhere as needed. To answer such a question, one can conduct tests on the "Lattice–Eratosthenes" algorithm, where the time increases parabolically with the increase in the sample volume. Thus, the estimates of all the prime numbers within the biggest number like 2^{256} , can take many months. You may also want to find primes modulo a field [3].

All these calculations will take much longer than simple iteration method to determine number within required boundaries with subsequent tests for its simplicity [4].

Probabilistic number simplicity tests have proven themselves to be quite good, which do not fully guarantee that the tested number is explicitly prime, but the probability increases with each iteration of the test. Thus, if the test is carried out several times, the probability of simplicity of the number increase. There are not many probabilistic algorithms. Many of them rely on Fermat's small theorem [5].

In this article, we present a probabilistic test for ease of numbers by Solovey –Strassen. Its main advantage with respect to the simple Fermat test was the ability to recognize Carmichael’s numbers as composite.

The purpose of this paper is to study the Solovey–Strassen test, the implementation of its algorithm in C# and the analysis of its arbitrariness. In the process of achieving this goal, the authors were formed and successfully solved the following tasks: research of the Solovey–Strassen test, identifying possible modifications of the test; analysis of test time in relation to other tests for simplicity [6].

II. THEORETICAL FOUNDATION

The Solovey–Strassen test for prostate detection in a number is based on Fermat's little theorem:

If p –simple number, and a – a_n integer that is mutually prime with n , so: $a^{n-1} \equiv 1 \pmod{n}$.

Therefore, for any a in the range $0 < a < n-1$.

The essence of the test is to test not over each number from the entire sequence, but over a random set of different random numbers k times.

In this case, to identify the Carmichael numbers, the properties of the Jacobi symbol are used. Number generated randomly during the test, and satisfying the equality:

$$\left(\frac{a}{n}\right) \equiv a^{\frac{n-1}{2}} \pmod{n}; \quad (1)$$

where $\left(\frac{a}{n}\right)$ –Jacobi symbol, called witness of simplicity n [7].

Depending on the values a and n , the Jacobi symbol can be 1 or –1. If at the end of the test witnesses, the prime number n has been discovered as much as iterations k , the number n is probably simple, with a probability $1-2^{-k}$.

Probabilistic tests are used in systems based on the factorization problem, such as RSA or Miller–Rabin scheme. The Solovey – Strassen test can be used wherever there is a need for a quick check of some number for simplicity.

However, in practice, the degree of reliability of the Solovey—Strassen test is not sufficient, instead the Miller—Rabin test is used, since its accuracy is much higher. Moreover, the combined use algorithms such as trial division and a test of Miller—Rabin, or consecutive testing, with proper choice of parameters we can obtain better results than when using each test individually [8].

III. TEST ALGORITHM FOR SOFTWARE IMPLEMENTATION

In: $n > 2$, an odd positive integer to be tested, k , a parameter that determines the accuracy of the test.

Out: compound, means that n is exactly compound, probably simple and means that n is likely to be $1-2^{-k}$ simple.

Cycle: $i=1,2,\dots,k$:

A =random integer between 2 and $n-1$, inclusive:

If $\text{GCD}(a, n) > 1$, **then:**

out, n is composite, and **break**.

If $a^{\frac{n-1}{2}} \not\equiv \left(\frac{a}{n}\right) \pmod{n}$; **then:**

out n is composite, and **break**.

Else Jump to the next iteration in the cycle.

Out: n —probably simple, with chance $1-2^{-k}$ [9].

In 2005 at the International conference "Informational Technologies" A.A. Balabanov, A.F. Agafonov, V.A. Ryku of the proposed upgraded test by Solovey—Strassen. The Solovey – Strassen test is based on the calculation of the Jacobi symbol, which takes time. The idea of improvement is that in accordance with the theorem of quadratic reciprocity of Gauss, to proceed to the calculation of the magnitude that is the inverse of the Jacobi symbol, which is a more simple procedure [10].

To speed up the search for GCD of numbers, we use a binary algorithm; this will reduce the time spent on finding a common divisor of numbers in the test algorithm. The main difference between the binary algorithm and the classical one is that it uses bitwise shifts of the number for the accelerated division by 2.

As fast algorithm for exponentiation will be its improved, recursive binary version, which will reduce time of operation [11].

IV. ALGORITHM OS JACOBI SYMBOL CALCULATION

1. If $\text{GCD}(a, b) \neq 1$, break and return 0.
2. $r := 1$.
3. If $a < 0$ so $a := -a$
If $b \pmod{4} = 3$ then $r := -r$
4. $t := 0$
While a – even
 $t := t + 1$
 $a := a / 2$
end of while;
If t – not even, then
If $b \pmod{8} = 3$ or 5 , then $r := -r$.
5. If $a \pmod{4} = b \pmod{4} = 3$, then $r := -r$.
 $c := a$; $a := b \pmod{c}$; $b := c$.
6. If $a \neq 0$, goto step 4, else break and return r .

V. ALGORITHM CODE ON C#

```
public static long Yacobi (long a, long b)
{
    int r = 1;
    while (a != 0)
```

```
{
    int t = 0;
    while ((a & 1) == 0)
    {
        t++;
        a >>= 1;
    }
    if ((t & 1) != 0)
    {
        long temp = b % 8;
        if (temp == 3 || temp == 5)
        {
            r = -r;
        }
    }
    long a4 = a % 4, b4 = b % 4;
    if (a4 == 3 && b4 == 3)
    {
        r = -r;
    }
    long c = a;
    a = b % c;
    b = c;
}
return r;
}
```

VI. ALGORITHM FOR FAST EXPONENTIATION

As mentioned earlier, to improve performance it is necessary to improve the speed of raising the number to a power modulo [12].

This is necessary to verify the basic assertion:

$$a^{\frac{n-1}{2}} \not\equiv \left(\frac{a}{n}\right) \pmod{n}.$$

To simplify writing, method that implements fast exponentiation will be called *POW*.

1. Input of the algorithm *POW* values are received:
 a – even, n – the desired power of a , Gf – modulo over a^n .

2. **If** $n=0$ **return** 1.

3. **If** n – not even, **return** result
(*call* *POW* ($a, n-1, Gf$)- a) modulo Gf .

4. **Else** a variable b is added,

4.1 **If** n divided by 4 without the rest, in that case the remainder of the division *POW* ($a, n/4, Gf$) on Gf is assigned to b .

4.1.1 b it is built in a square and the result is taken modulo Gf .

4.1.2 The result is then squared again.

4.1.3 **Return** remainder of division b and Gf .

4.2 **Else** b assign the remainder value of

$$\text{POW} \left(a, \frac{n}{4}, Gf \right) a^2 \text{ to } Gf.$$

Return remainder of division b^2 into Gf .

VII. IMPLEMENTATION OF ABOVE ALGORITHM IN C#

```
public static long binpow(long a, long n, long Gf)
{
    if (n == 0) return 1;
    if (n % 2 == 1) return (binpow(a, n - 1, Gf) * a) % Gf;
    else
```

```

{
    long b;
    if (n % 4 == 0)
    {
        b = binpow(a, n / 4, Gf) % Gf;
        b *= b; b *= (b % Gf);
        return b % Gf;
    }
    else
    {
        b = binpow(a, n / 2, Gf) % Gf;
        return (b * b) % Gf;
    }
}
}
}

```

VIII. C# CODE TO FIND THE LARGEST COMMON DIVISION

```

public static int BinaryGCD (int A, int B)
{
    int k = 1;
    while ((A != 0) && (B != 0))
    {
        while (((A & 1) == 0) && ((B & 1) == 0))
        {
            A >>= 1;
            B >>= 1;
            k <<= 1;
        }
        while ((A & 1) == 0) A >>= 1;
        while ((B & 1) == 0) B >>= 1;
        if (A >= B) A -= B; else B -= A;
    }
    return B * k;
}

```

IX. TEST RESULT

After writing the program in C#, tests were performed. On their basis, it can be concluded that this implementation is effective [13]. Table 1 shows the results of the program with numbers of different lengths in a cycle of 1000 repetitions.

TABLE I. TIME

Number X	X ≈250	X ≈10 ²	X ≈10 ⁵	X ≈10 ⁷	X ≈10 ⁹	X ≈10 ¹¹	X≈ 10 ¹²
Classic	0.144	0.243	1.522	2.320	3.854	5.216	7.562
Developed	0.257	0.351	0.629	0.921	1.237	1.695	2.350

Based on the results obtained, it can be noted that the efficiency of the implemented algorithm is worse on the numbers of smaller size, but much better when working with numbers from and higher. Based on the experiments performed, it can be concluded that the developed algorithm will be more effective in practice, since in many cases the main need for simplicity testing will be for large numbers.

X. CONCLUSION

A quick search of Prime numbers remains an important task today, since many cryptographic protocols use simple numbers to work. The algorithm implemented in the course of the program effectively copes with this task [14, 15].

The main innovation in the work done is the algorithm of fast erection to degree, which allowed reducing the calculation time.

Also, all the algorithms were implemented in their binary representation, which also reduced the cost of resources.

As work done result, it can be concluded that modified Solovoy-Strassen test is suitable and possible to use it effectively in larger cryptographic implementations.

REFERENCES

- [1] W. R. Alford, A. Granville, C. Pomerance (2004). «There are Infinitely Many Carmichael Numbers». *Annals of Mathematics* 139: 703-722. DOI: 10.2307/2118576.
- [2] R.M. Solovay and V.Strassen (1977, submitted in 1974). «A fast Monte-Carlo test for primality». *SIAM Journal on Computing* 6 (1): 84-85. DOI: 10.1137/0206006
- [3] D. Bernstein, N. Heninger, T. Lange “Fact Hacks: RSA factorization in the real world”, 2012. <https://www.iacr.org/archive/asiacrypt2008/53500477/53500477.pdf>
- [4] Zheltov S.A., “Effective Computing in the CUDA Architecture in Information Security Applications”. PhD dis. / Zheltov S.A. – M: IINTB RSUH, 2014 – 145 p.
- [5] File archive of students / Chuvash State Pedagogical University by name of I.Ya. Yakovlev / Chernikov A., Semenov I., Evaluation of software quality. Practisc.pdf – Access mode: <https://studfiles.net/preview/5850014/page:12> (circulation date on 15.02.2018).
- [6] Nasterenko A. Introduction to modern cryptography. Theoretical and numerical algo-rithms. — 2011. — pp. 79-90.
- [7]https://en.wikipedia.org/wiki/Solovay-Strassen_primality_test
- [8] “Integer Factorization Algorithms Connelly Barnes Department of Physics”, Oregon State University https://math.dartmouth.edu/archive/m56s14/public_html/proj/Howey0_proj.pdf
- [9] M. Dietzfelbinger. "Primality Testing in Polynomial Time, From Randomized Algorithms to "PRIMES Is in P"". *Lecture Notes in Computer Science*. 300.
- [10] <http://mathworld.wolfram.com/PocklingtonsTheorem.html>
- [11] R. Motwani; P. Raghavan (1995). *Randomized Algorithms*. Cambridge University Press. pp. 417-423. ISBN 0-521-47465-5.
- [12] Bai Sep Shi, “Polynomial Selection for the Number Field Sieve”, 2011. PhD Thesis of Australian National University <http://maths-people.anu.edu.au/~brent/pd/Bai-thesis.pdf>
- [13] N. Chaudhary “Metrics for Event Driven Software”, PhD. Scholar of Gautama Buddha University, India (IJACSA) *International Journal of Advanced Computer Science and Applications*, Vol.7, No.1, 2016. http://thesai.org/Volume7No1/Paper_12_-_Metrics_for_Event_Driven_Software.pdf
- [14] Kosyakov M.S., “Introduction to Distributed Computing”. St. Petersburg, 2014. – 155 p.
- [15] Ishmukhametov Sh.T., Rubtsova R.G., “Mathematical Bases of Information Security”, Electronic textbook for students of the Institute of Computational Mathematics and Informa-tion Technology – Kazan: Kazan Federal University, 2012. – 138 p.

Reliability Issues in the Parallel Dataflow Computing System

Nikolay Levchenko

*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
nick@ippm.ru

Anatoly Okunev

*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
oku@ippm.ru

Dmitry Zmejev

*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)*
Moscow, Russia
zmejevdn@ippm.ru

Abstract—When creating large computing systems (supercomputers), an important place is given to the reliability of these systems. The architecture of the parallel dataflow computing system (PDCS) that implements a dataflow computing model with a dynamically formed context and has peculiarities in the construction of nodes and blocks and algorithms for the operation of individual computational elements. Therefore, it requires new approaches to ensuring reliability compared to traditional systems. The article provides a comparison of traditional and dataflow computing models, briefly describes the PDCS architecture and its specificity. The basic features of the computing system that allow increasing the degree of reliability of computing facilities are listed. One of the variants of the system recovery mechanism in the event of a fault or failure of a computational core using the example of an error in the execution unit is given.

Keywords— *dataflow computing model, computing system reliability, local recovery tools*

I. INTRODUCTION

The concept of a dataflow computing model today is experiencing another attempt at its implementation. The first attempt was made in the 1970s, when the concept of dataflow computing was described in detail and the first prototypes based on it appeared. In those years, the element base was rapidly developing, making it possible to overcome new frontiers in the performance of traditional single-processor systems. This has become an obstacle to the development of a new concept, the main advantage of which is manifested when the program is parallelized into many processor elements. The second attempt was made at the junction of the 1980s and 1990s, when new programming methods began to emerge for parallelization and distribution of computations. This time, the obstacle was the already existing software, as well as the irrelevance of the problem of the computation parallelization [1-4].

The current crisis in the development of the element base and methods for distributing computations over a growing number of processors, which is clearly visible today, is forcing developers to revert to the experience of creating dataflow computing systems [5-6]. Until recently, most of the dataflow projects were limited to the development of specialized devices that accelerate the execution of a part of the program that is difficult to parallelize with traditional programming tools [7]. In the creation of a universal dataflow supercomputer, only a few are involved. In Russia, a suchlike project is underway at the Institute for Design Problems in Microelectronics of the Russian Academy of

Sciences – the Parallel Dataflow Computing System (PDCS) “Buran” [8].

New hardware and software solutions introduced into the PDCS architecture allow, in particular, the adjustment of delays in the communication network during data transmission; the localization of computations, minimizing long-distance data transfer between the cores; the reduction of the requirements for the amount of content addressable memory of the core, as well as the improvement of energy efficiency and fault tolerance of the system.

The goal of the article is to demonstrate approaches to ensuring the reliability of the dataflow architecture, a description of some schemes for implementing fault and failure recovery tools, as well as a description of the features of the dataflow computing model that can improve the reliability of such systems.

The reliability of the PDCS, which has in its composition hundreds of thousands, millions of cores, is one of the key problems solved in the design process. The concept of reliability includes many indicators. The objective of this article is to determine the main approaches to ensuring the fault tolerance of the system and the creation of automatic tools for restoring the operation of the computational core after a fault or failure.

II. COMPARISON OF TRADITIONAL AND DATAFLOW COMPUTING MODELS

Comparing the dataflow computing model with the traditional one, denote the basic computations control principles inherent in each model, regardless of the specific implementations. The control-flow concept is characterized by the sequential execution of instructions by the executive device. This sequence is formed by the algorithm described by the programmer in the form of a set of data processing operators and conditions for the transition between these operators. Execution of instructions consists in decoding instructions, loading operands from memory for processing, processing operands, and recording the result of processing into memory.

The dataflow concept is based on the initialization of computations by data readiness. The algorithm of the program represents not the sequence of instructions execution, but the conditions of data processing. These conditions are associative elements attached to each operand, and they are matched with other associative elements when this operand enters the memory. In case of matching with the associative element of the other operand, a packet of these

two data is formed, which is transmitted to the executive device, where the processing of this data is initialized. The result of packet processing is the formation of a new operand with a new associative element.

From a comparison of these two concepts, the following conclusions can be drawn:

- In the dataflow, the most commands are transferred from the executive device (processor) to the memory, where the matching of associative elements of the operands is performed according to certain rules. It is these rules are the command system of the dataflow system. Thus, the memory device is functionally and structurally significantly more complicated. Moreover, the role entrusted to the executive device implies data processing by one operator as well as the ability to execute a set of instructions - in this case, the second command system is introduced into the dataflow system.
- The programming paradigm for dataflow systems is radically different from the imperative, inherent in all modern programming languages and supported by traditional computing systems. The dataflow paradigm consists in forming the interaction between data through the nodes of their processing and is represented as a directed graph, where the arcs are data and the nodes are the operators of processing this data.
- Debugging imperative programs is simplified due to a deterministic sequence of instructions. Debugging programs in the form of a dataflow graph is complicated, since the order of activation of nodes is non-deterministic.
- Multitasking in dataflow systems is natural and does not require additional support by the operating system. This mode can be organized by adding an additional field to the associative element of each operand.
- The moment of the program end recognition in traditional systems is laid in the programming paradigm and is fixed when the special program end instruction is executed. In dataflow systems, this procedure is complicated by the non-determinism of the program execution process.
- The distribution of computations across several processors in the paradigm of imperative programming is a nontrivial and very laborious task, even when using specialized libraries and parallel programming systems such as Message Passing Interface (MPI). Moreover, the programming paradigm remains the same, and the mentioned software solutions only provide the transfer of data between processors. The whole concept of a dataflow computing model is based on data transfer, that is, the dataflow paradigm is inherently parallel. Parallelism is provided by the distribution of operands by hashing their associative elements. Moreover, since associative elements are numerical values, the distribution can be maintained at the hardware level and is not limited by the number of processor elements in the system - this number is a hashing

parameter and can take any value (within the available number of processors).

Based on this, it can be concluded that dataflow systems differ from traditional ones both in the nature of programming and in the interaction between the structural elements of the system.

All these features of dataflow systems (including the PDCS) undoubtedly affect the creation of tools that ensure the reliability of the nodes and blocks of the system.

III. DESCRIPTION OF THE PDCS COMPUTING MODEL AND ARCHITECTURE

The dataflow computing model with a dynamically formed context [9] is based on the activation of indivisible computational quantum by data readiness. A computational quantum is a program node, which, after being activated, is processed without interruption for additional external data; that, the program node operates only with data that came to its input.

In turn, the program for the parallel dataflow computing system (PDCS) is a set of descriptions of program nodes. Activation of any program node occurs only after all the necessary data elements (tokens) arrive at all its inputs. A token is a data structure containing a data, set of service fields and a key (index) that uniquely identifies the location of this data in the virtual address space of the task.

The input of the program node receives a packet containing data on which the program code is executed in this node. A packet is formed as a result of processing and defining ready-for-execution tokens in the matching processor. As a result of the program node operation, new values are calculated (solely on the basis of the values of the input data and fields of the packet key) and sent to other program nodes. Moreover, the key of the destination node is calculated directly in the same program node before sending data.

The PDCS, which is based on the dataflow computing model with a dynamically formed context, is a multi-core scalable computing system. Between the cores in the system, informational items are transferred in the form of tokens. Commutation between cores is based on the value of the core number generated by the hash block based on a parametrizable computation distribution function. The computational core includes a matching processor (MP) with content addressable memory of keys and token memory, execution units (EU), hash blocks, and an internal commutator of tokens.

Computational cores within a single crystal are organized into computational modules. The PDCS architecture is scalable and with an increase in the number of cores in the system, the drop in real performance on tasks with complexly organized data is much slower than in solving similar tasks on computing systems with classical architecture.

IV. FEATURES OF THE PDCS ARCHITECTURE THAT IMPLEMENTS DATAFLOW COMPUTING MODEL AND THEIR EFFECT ON RELIABILITY

The dataflow computing model with a dynamically formed context and its hardware implementation have special features that can increase the reliability of a

multiprocessor computing system. These features are taken into account when designing local recovery tools of the computational core and module, which allows increasing the fault tolerance of the system as a whole. These features are the following

A. The use of the hardware content addressable memory

The presence of hardware content addressable memory (CAM) allows the most natural way to activate the computational quanta and synchronize the computational process by data. The content addressable memory differs from direct accessible memory (used in traditional computing systems) by the way of addressing content, the ability to write to free space, as well as simultaneous viewing of all cells in search mode. When designing local recovery tools, standard solutions for recovering from a fault or failure can be used, but this is not enough, new approaches are needed.

B. The impersonality of execution units.

The matching processor provides in the PDCS architecture the synchronization of process of computation by data, through the matching of tokens. As a result of this matching, packets are formed. The packet has the property of independent execution, i.e. it can be processed on any free execution unit in the system. This fact simplifies local recovery after a fault on an individual EU and allows continuing working if one or several EUs fail. In addition, the manufacturability of crystals is improving.

C. The ability to begin computations before the completion of data formation.

One of the special features of the PDCS architecture is that the operation of the computing system can begin before the receipt of the complete set of input data. This concerns both the beginning of work on a specific task, and the transition to a new iteration in the course of computations within one task, thus making it possible to abandon global synchronization (if possible) during the transition from iteration to iteration. This feature is inextricably linked with the following one.

D. The use of the “scattering” paradigm and the principle of single assignment.

The use of the “scattering” paradigm implies that the generator of each new value knows who will need it, and independently provides the distribution to the right addresses. In this case, the recipient is left to passively “expect” the arrival of the data, without knowing anything about their source. This principle is best achieved through the use of the principle of single assignment, implemented using content addressable memory.

E. Non-deterministic computation process associated with the asynchronous execution of the program (organization of computations by data readiness).

The computational process in the PDCS is non-deterministic, unlike the result of the computation. It is this non-determinism at the hardware level that allows to extract the “implicit” parallelism embedded in the algorithm of the problem itself, which the programmer sometimes doesn’t see initially. Non-determinism itself is formed as a result of the organization of computations by data readiness, which, together with multiprocessing and possible problems of data

transmission over a communication network, ensures its existence.

All of the above features of the PDCS architecture and the dataflow computing model imply the development of original information collection algorithms for generating checkpoints, mechanisms for fixing faults and failures, as well as the use of new hardware solutions for creating local recovery tools for the computational core and the multicore computational module of the system.

V. LOCAL RECOVERY TOOLS

The standard operation in the parallel dataflow computing system is to continuously receive and send tokens both between computational cores and inside the core – between EU and MP (Fig. 1).

The general approach to implementing local recovery tools and the recovery process itself can be determined by the following steps:

- collection of information for the formation of local checkpoints (tokens arriving at the MP input; packets for execution on the EU; tokens leaving the EU);
- continuation of the program execution;
- localization of abnormal situation (in EU or MP) and setting control signals in the event of a fault or failure;
- fixation of the start of the fault handling process;
- start of the recovery mechanism after a fault or failure (fixation of the correctable and uncorrectable errors) from the local checkpoint in the EU or MP;
- proceeding to continuation of the program execution (if successfully overcoming a fault in the local element of the system or the failure of duplicate blocks of the computational core, such as the EU or hash block) or restarting the task (rebooting the system) with an uncorrectable error (with partial degradation of the system and its reconfiguration).

The creation of a local checkpoint (LCP) at the level of the computational core (or module) occurs in the absence of a global suspension of the computational process, and only the individual computational core (module) is suspended.

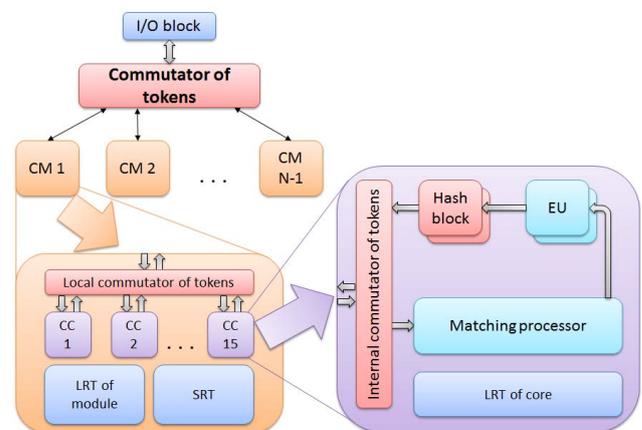


Fig. 1. Base architecture of the PDCS with local recovery tools (I/O block – input/output block; CM – computational module; CC – computational core; LRT – local recovery tool; SRT – system recovery tool; EU – execution unit)

When processing a fault in the EU, it takes a minimum of time to restore work — only to redirect part of the task to another EU.

To create LCP, there are two strategies: the first is to create a checkpoint after the end of a certain stage of computation on a given computational core, the second - at the start of the next stage of computation. In both cases, it is necessary to fix all tokens that have already been received or will be received from other cores.

Here is one of the variants of the system recovery mechanism in case of a fault or failure of a computational core using the example of an error in the EU. In the event of an abnormal situation in the EU, it is necessary to ensure the continuation of program execution. To do this, it is needed to fix the following information about the program node that was running at the time of the fault or failure in the EU:

- a packet that is executed in the EU, but the processing of which has not ended as a result of a fault or failure;
- all tokens that were generated in the EU before the interruption and were related to the packet that have been processed in the EU (or the number of these tokens).

Two buffers are added to the composition of the computational core: a token buffer, which is located at the output of the EU, and a packet register at the EU input.

Until the packet has been processed, the tokens from the buffer and the output of the EU are not transferred to the token commutator. In the normal, uninterrupted mode of the EU operation, the buffer starts to be freed only after the program node has completed its work (that is, when the packet has been processed). And only at this moment the tokens related to this node are transferred for further processing. When a fault occurs, the tokens from the buffer are destroyed and the packet is restarted. In the EU recovery mode, an eight-time restart of the packet that caused the fault occurs. If the restart is unsuccessful, the failure of the EU is fixed, after which this EU is excluded from the system operation. If the EU fails at the moment of the recovery mechanism initiation, it is closed to receive a packet from the MP, and the packets whose processing are not completed and cannot be continued on this EU are redirected through the packet commutator to another free EU.

VI. CONCLUSION

Ensuring the reliability of computing systems is currently receiving a great deal of attention, both in terms of fault-tolerant algorithms [10] and on the hardware side [11], this is especially true for computing systems of an exaflops level of performance.

The creation of a universal high-performance computing system that implements a dataflow computing model is carried out at IPPM RAS. The new architecture makes it possible to adjust data transfer delays, effectively localize computations, and reduce the requirements for the amount of content addressable memory used to match data contexts. At the same time, there is a real possibility to improve energy efficiency and fault tolerance of the system by hardware-software tools.

Reliability is one of the main indicators of the quality of high-performance computing systems. For systems such as

PDCS, new approaches to ensuring reliability are required. Moreover, the problem of increasing the reliability of such systems is aggravated by the presence of a large amount of parallel computing.

The dataflow computing model and the architecture that implements it have special features, using which the reliability of the computing system can be increased. For example, the impersonality of EUs allows the simplification of local recovery and the continuation of operation if one or more EUs fail. The use of the “gathering” paradigm and the principle of single assignment, as well as the ability to begin calculations before the data is completely formed, makes it often possible to abandon global synchronization during the transition to the next iteration.

These and other features require the development of original algorithms for collecting information for the formation of local checkpoints, methods of fixing faults and failures, as well as the introduction of new hardware into the composition of the computational core and module to automatically restore operation after a fault.

Based on the considered general approaches to ensuring reliability in dataflow computing systems, it is concluded that the development of local recovery tools has its own characteristics, allowing providing a high degree of fault tolerance required for such computing systems.

Currently the authors are working on a package of test programs that implement new algorithms for collecting information to create local checkpoints. Also, these programs will have to verify with a high degree of accuracy the hardware that recover the system after faults and failures. A series of experiments is planned to verify the hardware and software tools proposed for implementation in the PDCS "Buran", which are being developed to improve the reliability of the system.

REFERENCES

- [1] A. P. W. Böhm, “Dataflow and hybrid dataflow architecture summary,” in *Parallel computer systems*, Rebecca Koskela and Margaret Simmons (Eds.), ACM, New York, NY, USA, 1990, pp. 281-286. DOI=<http://dx.doi.org/10.1145/100215.100286>
- [2] B. Lee, A. R. Hurson, “Issues in Dataflow Computing,” *Advances in computers*, 1993, vol. 37, pp. 285-333.
- [3] B. Lee, A. R. Hurson, “Dataflow Architectures and Multithreading,” *Computer*, Aug 1994, vol. 27, no. 8, pp. 27-39.
- [4] J. Silc, B. Robic, T. Ungerer, “Asynchrony in parallel computing: From dataflow to multithreading,” *Parallel and Distributed Computing Practices*, 1998, vol. 1, no. 1, pp. 3-30.
- [5] Y. Birk, O. Mencer, “A Data Centric Perspective on Memory Placement,” in *Proceedings of the 2015 International Symposium on Memory Systems (MEMSYS '15)*, ACM, New York, NY, USA, 2015, pp. 39-42. DOI=<https://doi.org/10.1145/2818950.2818956>
- [6] G. Smaragdous, C. Davies, C. Strydis, I. Sourdis, C. Ciobanu, O. Mencer, and C. I. Zeeuw, “Real-Time Olivary Neuron Simulations on Dataflow Computing Machines,” in *Proceedings of the 29th International Conference on Supercomputing - Volume 8488 (ISC 2014)*, Julian Kunkel, Thomas Ludwig, and Hans Meuer (Eds.), Springer-Verlag New York, Inc., New York, NY, USA, 2014, pp. 487-497. DOI=http://dx.doi.org/10.1007/978-3-319-07518-1_34
- [7] T. Nowatzki, V. Gangadhar, N. Ardalani, and K. Sankaralingam, “Stream-Dataflow Acceleration,” in *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA '17)*, ACM, New York, NY, USA, 2017, pp. 416-429.
- [8] A. V. Klimov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, D. N. Zmejčev, “The PDCS "Buran" Operating Efficiency Improvement Ways,” in *Proceedings of IEEE EAST-WEST DESIGN & TEST SYMPOSIUM (EWDTS'2016)*, Yerevan, Armenia, October 14-17, 2016, pp. 323-326.

- [9] A. D. Ivannikov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, D. N. Zmejev, "Dataflow Computing Model – Perspectives, Advantages and Implementation," in Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2017), Novi Sad, Serbia, Sept 29 - Oct 2, 2017, pp. 187-190.
- [10] N. Abeyratne, H.-M. Chen, B. Oh, R. Dreslinski, C. Chakrabarti, and T. Mudge, "Checkpointing Exascale Memory Systems with Existing Memory Technologies," in Proceedings of the Second International Symposium on Memory Systems (MEMSYS '16), ACM, New York, NY, USA, 2016, pp. 18-29.
- [11] M. Kutlu, G. Agrawal, and O. Kurt, "Fault tolerant parallel data-intensive algorithms," in Proceedings of the 21st international symposium on High-Performance Parallel and Distributed Computing (HPDC '12), ACM, New York, NY, USA, 2012, pp. 133-134. DOI=<http://dx.doi.org/10.1145/2287076.2287099>

Modification and Optimization of Pollard's Factorization ρ -Method by Means of Recursive Algorithm of Number Calculation Factorization

Ivan A. Smirnov
Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
terran.doatk@mail.ru

Pavel V. Razumov
Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
razumov1996@inbox.ru

Nickolay V. Boldyrikhin
Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
boldyrikhin@mail.ru

Larissa V. Cherkesova
Mathematics and Computer Sciences Chair
Don State Technical University
Rostov-on-Don, Russia
chia2002@inbox.ru
ORCID 0000-0002-9392-3140

Yelena A. Revyakina
Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
revyelena@yandex.ru

Vitaliy M. Porksheyan
Applied Mathematics Chair
Rostov-on-Don, Russia
spu-46@donstu

Olga A. Safaryan
Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
safari_2006@mail.ru

Andrey G. Lobodenko
Information Systems and
Radioengineering
Don State Technical University
Rostov-on-Don, Russia
andrey@sssu.ru

Abstract— *Investigations of cryptographic algorithms for today are very actually in connection with cybernetic attacks threat and necessity of information protection at the enterprises of various levels including the strategic appointment. The project implementation of John Pollard's factorization ρ -method in the programming language C++ is presented, which works faster than the standard algorithm by 27%. It can facilitate greatly the deciphering operation and cryptographic analysis of various ciphers such as RSA cipher.*

Keywords— *factorization algorithm, Euclid, Pollard, algorithmic complexity, adjacency matrix, reachability matrix, McCabe metrics.*

I. INTRODUCTION

Cryptographic algorithms research is very important for today, in the connection with the cyberattacks threat (menace) and the necessity of protection of information at the enterprises of various levels, including the institutions strategic assignment.

Algorithm of encryption with public key RSA is still one of the most widely applied cryptographic algorithms. This asymmetric cryptographic algorithm is based on the factorization calculating of very big numbers. The RSA scheme uses as public, as private keys. The public key consists of open exponent (some number e) and module N , which is obtained by the product of simple integers P and Q [1]–[3].

After formation of public keys, a private key is generating. For this, it is necessary to calculate the Euler function $\varphi_n = (P-1) \times (Q-1)$ and element d , which is calculated by formula $d = e^{-1} \bmod \varphi^n$. With assistance of d element, it is possible to decipher the encrypted information: to factorize module N [4].

For decision the factorization problem, decomposition is used – scientific method that divides a large task into a series of smaller interconnected tasks. In this case, the factorization is the decomposition of the object of number N

into the product of two simple integers (numbers) that, when multiplied, will give the initial original object.

For example, the number 15 can be factorized into simple numbers 3 and 5, and the polynomial $x^2 - 81$ correspondingly at $(x-9) \cdot (x+9)$. Thanks to factorization operation, we can obtain the product of simpler objects.

However, in cryptographic algorithms very big numbers are used, which makes certain difficulties for factorization task. For such problems decision, many different crypto algorithms have been created, such as [5]:

- Shanks factorization cryptographic algorithm.
- Factorization algorithm of by means of elliptic curves.
- Miller–Rabin simplicity test.
- ρ -method of factorization – Pollard's algorithm, etc.

All of them, undoubtedly, have their merits and demerits (advantages and disadvantages)–in particular, one of the main shortcomings can be called a rather low operation speed, which, at modern cyber threats, is very considerably.

Let us consider the Pollard's algorithm.

The aim of our investigation is improvement of the ρ -method factorization Pollard's algorithm, whose modification is capable to increase the operation speed of the previously realized usual standard algorithm. The improved version of the algorithm was tested by three criteria of software reliability, including the McCabe metric. It has shown the better results, than usual standard algorithm.

For the factorization of integers, John Pollard invented his own algorithm in 1975, which was named ρ -algorithm by John Pollard in honour of its founder.

The foundation of this algorithm is the Robert Floyd's algorithm, invented by him in the late of 60–s of the XX-th century and which is effective for the searching of the cycle length in a sequence. Such mathematicians as John Pollard, Donald Knuth and others, have implemented a detailed analysis of this algorithm and proposed the several modifications and improvements of this algorithm [6].

The most efficient, at factorization of composite numbers with small multipliers, in decomposition is found ρ -algorithm.

A special feature of this algorithm is construction of numerical sequence in which, from a certain number n , elements of this sequence form a cycle (Figure 1).

This feature is represented in the form of the Greek letter “ ρ ”, which was the foundation for the naming of entire family of Pollard’s methods.

II. THEORETICAL FOUNDATION

In 1981, Richard Brent and John Pollard, using this algorithm, have found the smallest divisors of Fermat numbers $F_n = 2^{2^n} + 1$ at $5 \leq n \leq 13$.

So, $F_8 = 1238926361 \cdot 552897 \cdot p_{62}$, where p_{62} is prime (simple) number consisting of 62 decimal digits. In the “Cunningham project” framework, in 1925, the Pollard’s algorithm helped to find 19–digit divisor of number $2^{2386} + 1$.

Divisors of even larger numbers could be found also, but the creation and application of elliptical curves for factorization tasks made John Pollard’s algorithm less competitive and less in demand.

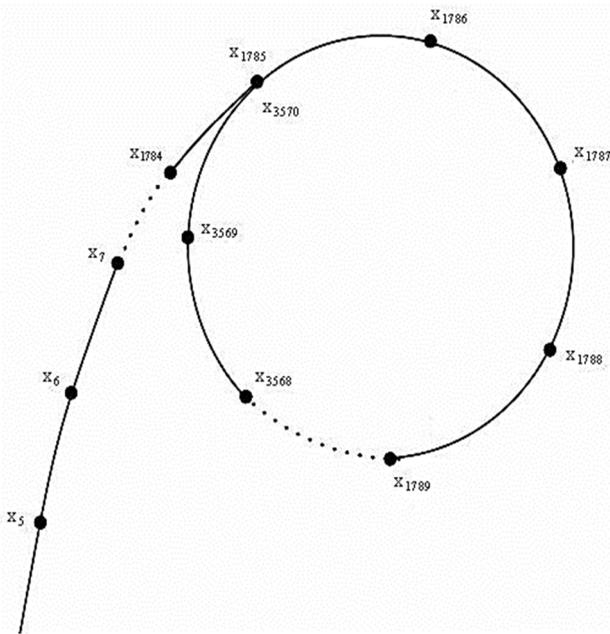


Fig. 1. The numerical sequence is looped starting from some number n .

Pollard’s ρ -method algorithm:

1. Let us consider the sequence of integers x_n , such that each next number $x_{i+1} = (x_i^2 - 1) \bmod N$, and $n = 0, 1, 2, \dots$. In this sequence x_0 – any small number.

2. At each step we will calculate the value $d = \text{GCD}(n, |x_i - x_j|)$, where $j < i$.

3. If $d \neq 1$, then the calculations are finished. The founded number d is the divisor of number n . If n/d is not prime (simple) number, then the procedure can be continued by taking the number n/d instead of number n [7].

Instead of the function $F(x) = (x^2 - 1) \bmod n$ for the calculation x_{n+1} , we can take another polynomial, for example, $x^2 + 1$, or some other polynomial of the second degree:

$$F(x) = ax^2 + bx + c. \quad (1)$$

The main drawback of this method is the necessity of additional memory allocation for storing the previous values of x_j .

Let us note that if $(x_i - x_j) \equiv 0 \pmod{p}$ then $(f(x_j) - f(x_i)) \equiv 0 \pmod{p}$ then if the pair (x_i, x_j) gives us solution, then some decision will give any other pair [7], [8]:

$$(x_i + k, x_j + k). \quad (2)$$

In this connection, there is no necessity to check all pairs (x_i, x_j) , but there is the possibility to restrict oneself only to pairs of the form (x_i, x_j) , where $j = 2^k$, and k passes a set of consecutive values 1, 2, 3, ..., and i will take values from the gap $[2^k + 1; 2^{k+1}]$.

Another modernization of the Pollard’s ρ -method was created by Floyd, according to which the value of y varies at the each step according to the formula:

$$y = F_2(y) = F(F(y)). \quad (3)$$

Therefore, in the step i , values, $x_i = F^i(x_0)$, $y_i = x_{2i} = F^{2i}(x_0)$, and greatest common divisor (GCD) will be found at this step, it is calculated between values n and $y - x$ [9].

Substantiation of the Pollard’s ρ -method

Let us consider this method and calculate its labor intensiveness. Such estimation is based on the well-known theorem by name of “birthday paradox”.

The theorem. Let $\lambda > 0$. For arbitrary selection from $l+1$ element, each of which is less than the number q , where $l = \sqrt{2\lambda q}$, the probability p that two elements are equal, will satisfy the inequality $p > 1 - e^{-\lambda}$.

The probability $p = 0,5$ in the “birthday paradox” is obtained at $\lambda \approx 0,69$.

Let us assume that the order of the elements $\{u_n\}$ consists of the differences $|x_i - x_j|$ that are checking in the algorithm’s operation process. We will establish some new sequence $\{z_n\}$, where $z_n = u_n \bmod q$ and q is the smaller of the divisors of the number n . All elements of the sequence $\{z_n\}$ are smaller than \sqrt{n} . If we regard $\{z_n\}$ as random sequence of numbers less than q , then, correspondingly the *twins paradox*, the probability that in the number of the first $l+1$ of its terms will contain two identical ones will exceed $1/2$ for $\lambda \approx 0,69$, then l should not be less than $\sqrt{2\lambda q} \approx \sqrt{1,4q} \approx 1,18\sqrt{q}$.

If $z_i = z_j$ then $x_i - x_j \equiv 0 \pmod{q} \rightarrow x_i - x_j = kq$ for some $k \in \mathbb{Z}$.

If $x_i = x_j$, that is having high probability, then the required divisor q of the number n will be found as $\text{GCD}(n, x_i - x_j)$. In the view of the fact that $\sqrt{q} \leq n^{1/4}$, then with probability greater than 0,5 the divisor n can be found for 1,18 $n^{1/4}$ iterations.

As we see, the Pollard's ρ -method is a probabilistic method that allows us to find the nontrivial divisor q of the number n for $O(q^{1/2}) \leq O(n^{1/4})$ iterations. The complexity of nontrivial divisor finding in this method depends only on the size of this divisor, and not on the size of the number n . In this connection, the Pollard's ρ -method is used in those cases when other factorization methods, which depend on the size of n , are ineffective.

In other cases, the sequence $\{y_n\}$ will be looped (i.e. at the certain step t appears $x_t = x_0$, then the sequence repeats), then we need to replace the current element x_0 or the polynomial $F(x)$ by some other element [4], [7].

Software implementation and analysis of the algorithm

Let us represent this algorithm in the form of a scheme, and calculate its structural complexity. The usual original Pollard's algorithm is shown in Figure2.

```

int  $\rho$ -Pollard (int n)
{ int x = random (1, n-2);
  int y = 1; int i = 0; int stage = 2;
  while(H.O.D. (n, abs(x - y)) = 1)
  {
    if (i == stage) {
      y = x;
      stage = stage*2; }
    x=x * x + 1(modn);
    i=i + 1;
  }
  return H.O.D. (n, abs(x - y)); }

```

Fig. 2. Original Pollard's algorithm.

III. RESULTS AND DISCUSSIONS

For quick finding the GCD, it makes sense to take advantage of *Euclid's binary algorithm*. Its preference over the usual algorithm consists in using of *bitwise shifts*, which, according to various estimations, have advantage in operation speed up to 30%. Therefore, it is expected that at using the binary Euclid's algorithm, we will get factorization of numbers on 30% faster than it would be obtained at application the usual standard algorithm.

Modification of ρ -method of Pollard's algorithm. The modification of algorithm is based on *recursive method* for calculating of the number factorization. Such modified algorithm works on 27% faster than usual original Pollard's algorithm based on iterations.

For qualitative analysis and comparison of two algorithms, we will consider their flowcharts (algorithm schemes) presented in Figures 3 and 4.

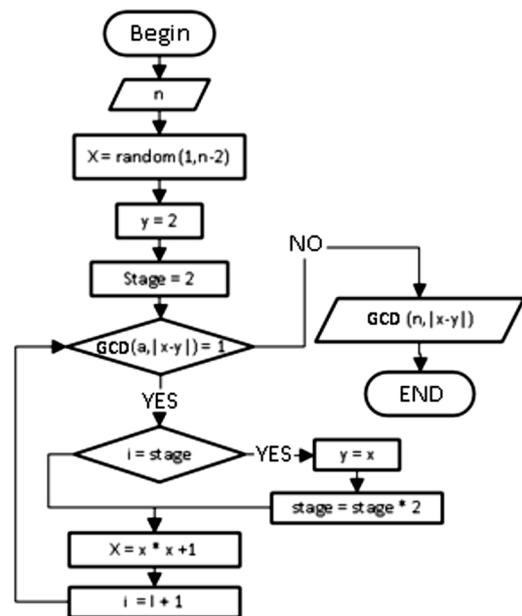


Fig. 3. Usual standard Pollard's iterative algorithm.

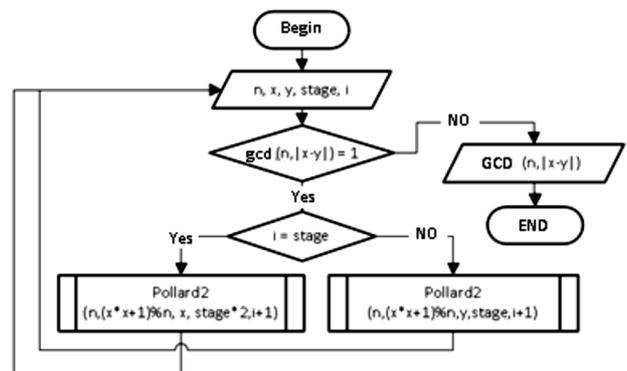


Fig. 4. Modified Pollard's recursive algorithm.

Table I demonstrates the program code of these two algorithms.

TABLE I. THE PROGRAM CODE OF TWO ALGORITHMS

Iterative Pollard's algorithm	Modified recursive Pollard's algorithm
<i>Softwarecode</i>	
<pre> int Pollard3(int n, int x1) { srand(time(NULL)); int y = 1; int i = 0; int stage=2; while (gcd_bynary(n, abs(x1-y))==1) { if (i == stage) { y = x1; stage = stage * 2; } x1 = (x1 * x1 + 1) % n; i = i + 1; } Return gcd_bynary (n, abs(x1 - y)); } </pre>	<pre> int Pollard2(int n, int x, int y, int stage, int i) { if (gcd_bynary(n, abs(x - y)) == 1) { if (i == stage) { return Pollard2(n, (x*x+1)%n, x, stage*2, i+1); } return Pollard2(n, (x*x+1)%n, y, stage, i+1); } Else Return gcd_bynary(n, abs(x - y)); } </pre>

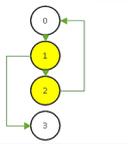
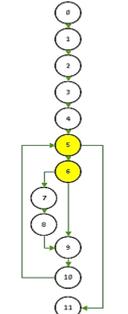
Now let us evaluate the structural complexity of the software using three criteria to ensure the advantage of the developed method before the standard implementation.

Criterion 1: according to this criterion, the graph of control flow by program must be checked for the smallest set of routes passing through each branch operator along each arc [2].

Passage for each route occurs no more than once, because repeated checking of arcs is considered as excessive (redundant).

During the check, it is ensured that all control transmissions are executed among the program operators and each operator, at least once. It should be noted that there are algorithms that allow improving the process of the minimum set of routes obtaining by this criterion [5], [10]. Table II demonstrates the graphs and program codes of these two algorithms.

TABLE II. COMPARISON OF ITERATIVE AND RECURSIVE POLLARD'S ALGORITHMS

Created by authors Pollard's recursive algorithm	
	<pre>int Pollard2(intn, intx, inty, intstage, inti) //0 { if (gcd_byary(n, abs(x - y)) == 1) //1 { if (i == stage) //2 { return Pollard2(n, (x*x+1)%n, x, stage*2, i+1); } return Pollard2(n, (x*x+1)%n, y, stage, i+1); } Else Return gcd_byary(n, abs(x - y)); //3</pre>
Usual standard iterative Pollard's algorithm	
	<pre>int Pollard3(intn, intx1) //0 { srand(time(NULL)); intx1 = rand() % n - 2 + 1; //1 int y = 1; //2 inti = 0; //3 int stage=2; //4 while (gcd_byary(n, abs(x1-y))==1) //5 { if (i == stage) //6 { y = x1; //7 stage = stage * 2; //8; x1 = (x1 * x1 + 1) % n; //9 i = i + 1; //10; Return gcd_byary(n, abs(x1 - y)); //11}</pre>

Estimation of algorithmic complexity

We will define minimal set of routes passing through any branching operator and for each arc, as presented in the Table III.

TABLE III. DETERMINING THE MINIMAL SET OF ROUTES

int Pollard3 (int n, int x ₁)	int Pollard2(int n, int x, int y, int stage, int i)
m ₁ : 0-1-2-3-4-5-11; p ₁ = 1	m ₁ : 0-1-3; p ₁ = 1;
m ₂ : 0-1-2-3-4-5-6-7-8-9-10-5; p ₂ = 3	m ₂ : 0-1-2-0-1-3; p ₂ = 2;
m ₃ : 0-1-2-3-4-5-6-9-10-5; p ₃ = 3	

In accordance with the *first criterion* of algorithmic complexity estimation, the required number of routes for the standard realization of ρ-method Pollard is equal to 3, and for the developed by authors algorithm, is equal 2.

The complexity level determines the number of branching vertices in the graphs:

$$S_1 = p_1 + p_2 = 1 + 2 = 3 - \text{int Pollard2}(\text{int } n, \text{int } x, \text{int } y, \text{int } \text{stage}, \text{int } i);$$

$$S_2 = p_1 + p_2 + p_3 = 1 + 3 + 3 = 7 - \text{int Pollard3}(\text{int } n, \text{int } x_1).$$

Criterion2: this criterion is based on an analysis of the *baseline routes* in the program that are generated and estimated on the base of cyclomatic number, determined with assistance of the graph of program control flow. For each linearly independent cycle and acyclic section of the program, we define the number of checks [5].

The number of checks is determined by the cyclomatic number of the graph, which is defined by the following relationship [10]:

$$Z = n_B + 1 = 1 + 1 = 2 - \text{int Pollard2}(\text{int } n, \text{int } x, \text{int } y, \text{int } \text{stage}, \text{int } i);$$

$$Z = n_B + 1 = 2 + 1 = 2 - \text{int Pollard3}(\text{int } n, \text{int } x_1).$$

where n_B is the number of branching vertices. Next, it is necessary to select the routes on the specified preset graph. The results are demonstrated in the Table IV.

TABLE IV. CALCULATION OF ROUTES ON GIVEN GRAPHS

intPollard3 (intn, intx1)	intPollard2 (int n, int x, int y, int stage, int i)
<i>Acyclic</i>	
m ₁ : 0-1-2-3-4-5-11; p ₁ =1;	m ₁ : 0-1-3; p ₁ = 1;
<i>Cyclic</i>	
m ₂ : 5-6-7-8-9-10; p ₂ =2;	m ₂ : 1-2; p ₂ =2.

The testing of the program on the specified routes will allow checking all the branching operators and statements of the program. The following relation determines the metric of structural complexity:

$$Z = n_B + 1 = 2 + 1 = 2 - \text{int Pollard3}(\text{int } n, \text{int } x_1);$$

$$S_2 = \rho_1 + \rho_2 = 1 + 2 = 3;$$

$$Z = n_B + 1 = 1 + 1 = 2 - \text{int Pollard2}(\text{int } n, \text{int } x, \text{int } y, \text{int } \text{stage}, \text{int } i);$$

$$S_2 = \rho_1 + \rho_2 = 1 + 2 = 3.$$

The next step is the construction of *adjacency matrixes*, with assistance of which the graph analysis is constructed. These matrixes contain the information about the structure of the program being tested [11].

Adjacency matrix is square matrix in corresponding cells of which contain units, if there is corresponding arc in control flow graph of the program. In another cases, this cell does not fill out. Adjacent matrixes of two algorithms are shown in Tables V and VI.

TABLE V. ADJACENCY MATRIX OF THE ITERATIVE ALGORITHM intPollard3(intn, intx1)

	0	1	2	3	4	5	6	7	8	9	10	11
0												
1	1											
2		1										
3			1									
4				1								
5					1							
6						1						
7							1					
8								1				
9									1			
10										1		
11						1					1	

TABLE VI. ADJACENCY MATRIX OF RECURSIVE ALGORITHM intPollard2(intn, intx, inty, intstage, inti)

	0	1	2	3
0			1	
1	1			
2		1		
3			1	

After the construction of the adjacency matrixes, the *reachability matrix* is constructed. In the cells of this matrix, the units are located at the position corresponding to the arc (i, j) . With assistance of computer tools, such a matrix can be obtained by raising the previous *adjacency matrix* in degree, whose value is equal to the number of vertices without the last one in initial graph of control flow.

Using the *reachability matrix*, it is possible to select out the cycles, noting diagonal elements, equal to unity (1), and identical rows. The reachability matrixes of two algorithms are shown in the Tables VII and VIII.

TABLE VII. REACHABILITY MATRIX OF THE ITERATIVE ALGORITHM $intPollard3(intn, intx1)$

	0	1	2	3	4	5	6	7	8	9	10	11
0												
1	1											
2	1	1										
3	1	1	1									
4	1	1	1	1								
5	1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1

TABLE VIII. REACHABILITY MATRIX OF RECURSIVE ALGORITHM $intPollard2(intn, intx, inty, intstage, inti)$

	0	1	2	3
0	1	1	1	
1	1	1	1	
2	1	1	1	
3	1	1	1	

Criterion 3: with assistance of this criterion, a complete composition of basic structures of the program control flow graph is formed, and each cyclic and acyclic route of the initial program graph, achievable from all these routes, is analyzed. According to this criterion, all really possible control routes were identified [5]. Their calculation is presented in the Table IX.

TABLE IX. CALCULATION OF ALL POSSIBLE CONTROL ROUTES

int Pollard3 (int n, int x ₁)	int Pollard2 (int n, int x, int y, int stage, int i)
$m_1: 0-1-2-3-4-5-6-7-8-9-10-5-11;$ $p_1 = 3;$ $m_2: 0-1-2-3-4-5-6-9-10-5-11;$ $p_2 = 3;$ $m_3: 0-1-2-3-4-5-6-7-8-9-10-5-$ $6-9-10-5-11; p_3 = 5;$ $m_4: 0-1-2-3-4-5-6-9-10-5-6-7-$ $8-9-10-5-11; p_4 = 5.$	$m_1: 0-1-2-0-1-3; p=2;$ $m_2: 0-1-3; p=1.$

The estimation of the structural complexity of the program for the corresponding algorithm has the following form:

$$int\ Pollard3(int\ n, int\ x_1): S_3 = p_1 + p_2 + p_3 + p_4 = 3 + 3 + 5 + 5 = 16 ;$$

$$int\ Pollard2(int\ n, int\ x, int\ y, int\ stage, int\ i): S_3 = p_1 + p_2 = 1 + 2 = 3 .$$

Conclusion on evaluation of structural complexity $intPollard3(intn, intx1)$: based on the obtained results of calculating the metrics of the function by the three criteria for selecting routes, we can conclude that the usual standard function of the Pollard ρ -method has the higher algorithmic complexity than the one developed by authors, and the number of conditional operators in the standard function will require at least two or three testing versions of the initial original data.

The same conclusion on evaluation or estimation of structural complexity $intPollard2(intn, intx, inty, intstage, inti)$: based on the obtained results of calculation of the function metrics by the three criteria for route allocation, we can conclude that the developed function has the lower algorithmic complexity compared to the previous one,

because one conditional operator is used and for which it suffices to check from one to two test variants of the initial original data [12].

TABLE X. THE RESULT OF THE ALGORITHMS OPERATION OF ON DIFFERENT PROCESSORS

Type of processor	Numbers				Average CPU efficiency	Total average efficiency
	11	22	456789 I	4567884		
Intel® Core(TM)30% i5-4200U CPU 2.3GHz	1832 / 2597 (30 %)	911 / 1235 (26 %)	20098 / 22743 (12 %)	827 / 1177 (30%)	27 %	27,6 %
Intel® Core(TM) i3-2330M 2.20 GHz	2845 / 4960 (43 %)	3429 / 6641 (49 %)	27443 / 44848 (39%)	4090 / 8468 (52%)	45,75 %	
AMD Phenom (tm) II Quad-Core Processor 1.80 GHz	143209 / 3485 (8 %)	3057 / 3545 (14%)	11242 / 11735 (5%)	3057 / 3545 (14%)	10,25 %	

Metrics of McCabe. This metric allows estimating the structural complexity of software tools built on the foundation of analysis of the flow of control from one operator to another one.

This will help to take into account the logic of constructing the program when evaluating of its complexity [12], [13].

In accordance with the control graph, the number of arcs is m ; the number of vertices is n . Then the cyclomatic McCabe number is:

$$intPollard3(intn, intx1): Z = m - n + 2 = 13 - 12 + 2 = 3 ;$$

$$intPollard2(intn, intx, inty, intstage, inti):$$

$$Z = m - n + 2 = 4 - 4 + 2 = 2 .$$

Testing. After the algorithm implementation, the independent testing was carried out on three (3) various computers with different processor's frequencies [11].

The results of testing are demonstrated in the Table X. Table 10 shows the execution time of the high-frequency counter functions. The numerator of the table cells shows the received data on the modified by the authors recursive method. In the denominator, there are the received data by the usual standard iteration Pollard's method. The calculated percentage under the counter demonstrates the efficiency of the first method in comparison with the second [14], [15].

IV. CONCLUSION

The aim of the research was improving of standard classical algorithm of Pollard's factorization ρ -method. Its modification, optimization and modernization significantly increases the performance of the standard algorithm. The peculiarity of the author's development, in comparison with classical standard algorithm, is the rapid finding of the greatest

common divisor GCD using Euclid's binary algorithm. Its advantage over the classical algorithm of Pollard's factorization ρ -method is using of *bitwise shifts*, which have advantage in speed up to 30%. Therefore, when using the binary Euclid algorithm, the factorization of numbers is 30% faster than in conventional classical algorithm.

The authors developed modification of algorithm based on the recursive method of number factorization counting, works on 27% faster than classical Pollard algorithm using iterations. Improved version of algorithm was tested by three-reliability criteria software, including McCabe metrics [13], and shown significant improvement in results.

Proceeding from this, developed by the authors and implemented through recursion the algorithm of ρ -method factorization by Pollard is more reliable and faster than the implementation of the usual standard iterative algorithm, which has been experimentally confirmed in practice by the method of computational experiment.

The application of this modified factorization algorithm in the RSA type of ciphers will allow to complicate the cryptographic analysis, which can be based on analysis of calculation of operation time in OpenSSL protocol, as well as cybernetic attacks based on the analysis of electric power expenditures [14], [15].

Using this modernized by authors algorithm of Pollard's ρ -method factorization will allow to protect the data from some kinds of cybernetic attacks, which promotes and contributes to the problem of information protection on the enterprises and institution of various levels, including strategic purpose.

REFERENCES

[1] Y. Song Yan, "Cryptanalytic Attacks on RSA", Proceedings of University of Bedfordshire, UK and Massachusetts Institute of Technology, USA, 2008.

[2] P. Kocher, "Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems", Advances in Cryptology Crypto '96, Springer-Verlag, LNCS 1109, 1996, Pp.104-113.

[3] D. Bernstein, N. Heninger, T. Lange, "Fact Hacks: RSA factorization in the real world", 2012. <https://www.iacr.org/archive/asiacrypt2008/53500477/53500477.pdf>

[4] Sh. Ishmukhametov, R. Rubtsova, "Mathematical Bases of Information Security", Electronic textbook for students of the Institute of Computational Mathematics and Information Technology – Kazan: Kazan Federal University, 2012. – 138 p. (In Russian).

[5] File archive of students / Chuvash State Pedagogical University by name of I. Yakovlev / A. Chernikov, I. Semenov, "Evaluation of software quality". Practisc.pdf-Access mode: <https://studfiles.net/preview/5850014/page:12/> (circulation date on 15.02.2018).

Softwarecode	
	<pre> int Pollard2(intn, intx, inty, intstage, inti) {if (gcd_bynary(n, abs(x - y)) == 1) { if (i == stage) { return Pollard2(n,(x*x+1)% n, x, stage*2, i+1); } return Pollard2(n,(x*x+1)%n, y, stage, i+1); } Else Return gcd_bynary(n, abs(x - y)); } </pre>

[6] Wikipedia/Ro-Algorithm of Pollard-Access mode: https://ru.wikipedia.org/wiki/Ro-Pollard_algorithm (circulation date 12.02.2018)

[7] <https://www.geeksforgeeks.org/pollards-rho-algorithm-prime-factorization/>

[8] "Integer Factorization Algorithms Connelly Barnes Department of Physics", Oregon State University https://math.dartmouth.edu/archive/m56s14/public_html/proj/Howey0_proj.pdf

[9] J. Hee Cheon, J. Hong, M. Kim "Speeding Up the Pollard Rho Method on Prime Fields"/ Department of Mathem. Sciences, Seoul National University, Seoul 151-747, Korea, 1997. snu.ac.kr

[10] P. Montgomery "Speeding the Pollard and Elliptic Curve Methods of Factorization", Mathematics of Computation. V.48, N.177, Pp. 243-264 <http://www.connellybarnes.com/documents/factoring.pdf>

[11] E. Howey, "Primality Testing and Factorization Methods", May 27/ 2014/ <https://pdfs.semanticscholar.org/7fd4/4eb2df7b39716e984d548c28f51d9dc6b6bbb.pdf>

[12] Bai Sep Shi, "Polynomial Selection for the Number Field Sieve", 2011. PhD Thesis of Australian National University <http://maths-people.anu.edu.au/~brent/pd/Bai-thesis.pdf>

[13] N. Chaudhary "Metrics for Event Driven Software", PhD. Scholar of Gautama Buddha University, India (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, 2016. http://thesai.org/Volume7No1/Paper_12-Metrics_for_Event_Driven_Software.pdf

[14] M. Kosyakov, "Introduction to Distributed Computing". St. Petersburg, 2014. – 155 p.

[15] S. Zheltov "Effective Computing in the CUDA Architecture in the Information Security Applications". PhD dis. / Zheltov S.A. – M: IINTB RSUH, 2014 – 145 p.

Deriving Low Power Test Sequences Detecting Robust Testable PDFs

A. Matrosova

*Institute of Applied Mathematics and
Computer Science
Tomsk state university
Tomsk, Russia
mau11@yandex.ru*

V. Andreeva

*Institute of Applied Mathematics and
Computer Science
Tomsk state university
Tomsk, Russia
avv.21@mail.ru*

V. Tychinskiy

*Institute of Applied Mathematics and
Computer Science
Tomsk state university
Tomsk, Russia
tvz.041@gmail.com*

Abstract—New approach to deriving low power test sequences that detects robust testable PDFs in logical circuits is suggested. Decreasing power consumption is provided by decreasing the number of switches during testing and cutting the sequence length. The approach is based on finding all test pairs consisting of neighbor Boolean vectors for a circuit path. The test pairs are compactly represented by the proper ROBDD. Each pair of neighbor Boolean vectors generates three neighbor Boolean vectors that detect robust testable PDF for both rising and falling transitions. Current approaches are oriented to finding if only one test pair for rising (falling) transition of a circuit path. Applying all test pairs and using intersections of ROBDDs representing these pairs we provide addition facilities to cut power consumption of the test sequences and their lengths. Different algorithms of deriving the test sequences are suggested. Experimental results demonstrate high quality of the test sequences.

Keywords—combinational and sequential circuits, reduced ordered binary decision diagrams (ROBDDs), robust testable path delay faults (robust testable PDFs)

I. INTRODUCTION

Power consumption of logical circuits may be essentially increased during logical circuit testing in the frame of scan techniques as compared with conventional circuit functioning. It is necessary to decrease it. Here we consider testing of robust testable Path Delay Faults (PDFs). It is known that such testing allows identifying the fault path and, if possible, removing its delay in order to increase circuit performance. We study facility of decreasing of power consumption taking into consideration circuit switching activity and cutting the sequence length. Our approach is oriented to Hamming distance reduction between neighbor test patterns. The approach may be applied in the frame of Random Access Scan (RAS) techniques. We suggest finding all test pairs consisting of neighbor Boolean vectors for each path from the given set. The test pairs (v_1, v_2) are compactly represented by the proper ROBDD. Current approaches [1-6] are oriented to finding if only one test pair for rising (falling) transition of a circuit path. Applying all test pairs of neighbor Boolean vectors and using intersection of ROBDDs representing test pairs of different paths we provide addition facilities of cutting power consumption of test sequences and decreasing their lengths.

In [7] the method of deriving all test pairs of neighbor Boolean vectors detecting robust testable PDFs for a circuit path based on operations on ROBDDs is suggested. The operations are executed on ROBDDs derived from fragments of a combinational circuit (the combinational part of a sequential circuit). It allows compact representing all test

pairs that detect robust testable PDFs of the path by the proper ROBDD R_{rob} . Each test pair presented by ROBDD R_{rob} originates three Boolean vectors: either (v_1, v_2, v_1) or (v_2, v_1, v_2) that detect robust testable PDFs both for rising and falling transitions of the corresponding path [7].

In section II the problem statement is discussed, in section III facilities of ROBDD R_{rob} intersections for different paths of a circuit are considered, in section IV the algorithms of extracting a cube from ROBDD as much as possible close to the given cube is described, in Section V algorithms of deriving test sequences are presented and experimental results are discussed.

II. PROBLEM STATEMENT

We have a set of ROBDDs: $R_{rob1}, \dots, R_{robL}$, corresponding to L paths of circuit C . In each ROBDD the variable correlated to the beginning of the path is absent. It is necessary to get the sequence of Boolean vectors that detects L robust testable PDFs. These vectors depend on n input variables of circuit C . We may use intersections of ROBDDs of different paths in order to cut the sequence length and get the sequence fragments with minimal power consumption. The fragments have the length 5 and more. Note that transition from one fragment to another, for example, from triple $v_1(i)v_2(i)v_1(i)$ of Boolean vectors (the shortest i -th fragment) to the next triple $v_1(i+1)v_2(i+1)v_1(i+1)$ (the shortest $(i+1)$ -th fragment) of the test sequence is connected with changing vector $v_1(i)$ for vector $v_1(i+1)$. Implementing this procedure we try to provide as less as possible Hamming distance between these two Boolean vectors using algorithm described below.

III. ROBDD R_{ROB} PROPERTIES

Remind that ROBDDs represent Disjoint Sum of Products (DSoP). Each product (cube) of DSoP is generated by the path connecting the ROBDD root with its 1 terminal node.

Two ROBDDs are orthogonal if their DSoPs don't intersect.

Two Boolean vectors are neighbor, if they differ by values of only one variable.

First consider two ROBDDs R_{rob1}, R_{rob2} corresponding to the different paths of circuit C that of which beginnings are marked by the same input variable x_i . Denote these ROBDDs as $R_1(x_i), R_2(x_i)$.

Theorem 1. An intersection of ROBDDs $R_1(x_i), R_2(x_i)$ is empty:

$$R_1(x_i) \& R_2(x_i) = \emptyset.$$

Proof. Admit opposite: an intersection of ROBDDs $R_1(x_i), R_2(x_i)$ is nonempty. It means that these ROBDDs originate even of a pair of products k_1, k_2 , one from each ROBDD, that provide nonempty intersection: $k_1 \& k_2 \neq \emptyset$. Remind that k_1 (k_2) represents the proper set of minimal cubes of rank $(n-1)$ that of which contains vectors of test pair (v_1, v_2) for the corresponding path of circuit C . Therefore, an intersection even of two cubes of rank $(n-1)$ (one from each products k_1, k_2) is nonempty. Note that the result of their intersections can be only b_p test pattern [9] but in this case their a_p test patterns has to coincide. The last is impossible: different paths have different a_p test patterns. We got a contradiction. The theorem is proved.

Now consider two ROBDDs R_{rob1}, R_{rob2} corresponding to the different paths of circuit C that of which beginnings are marked by different input variables x_i, x_j . Denote these ROBDDs as $R_1(x_i), R_2(x_j)$.

Theorem 2. An intersection of ROBDDs $R_1(x_i), R_2(x_j)$ may be nonempty:

$$R_1(x_i) \& R_2(x_j) \neq \emptyset.$$

Proof. Let ROBDDs $R_1(x_i), R_2(x_j)$ originate a pair of products k_1, k_2 , one from each ROBDD, that provides non empty intersection: $k_1 \& k_2 \neq \emptyset$. Consider possible results of their intersections on variables x_i, x_j . First, results from a set $\{00, 01, 10, 11\}$ of Boolean vectors on these variables. Take one result, for example, vector 11. This vector corresponds to Boolean vector β on n variables (in vector β variables x_i, x_j have values 11). Let vector β turns circuit C into 1 (0). Using β we derive test pairs for each of considered paths of circuit C . If vector β turns circuit C into 1, then β is a_p test pattern [7] (vector v_1) for the path, conforming, for example, to variable x_i . Then for the same path test pattern b_p contains vector 01 on variables x_i, x_j . Values of the rest variables of vectors β, b_p are the same. For simplicity we further represent Boolean vectors by only their components corresponding to variables x_i, x_j taking into consideration that the rest values are the same. The sequence from three vectors v_1, v_2, v_1 (11, 01, 11) ends in vector 11 that is also a_p test pattern (vector v_1) for the path conforming to variable x_j . Then Boolean vector 10 represents b_p test pattern (vector v_2) for the path conforming to variable x_j . Thus we may form the sequence from 5 vectors: (11, 01, 11, 10, 11). This sequence first detects robust PDF of rising and falling transitions for the path corresponding to variable x_i and then rising and falling transitions for the path corresponding to variable x_j . Taking into considerations properties of a_p, b_p test patterns we conclude that if an intersection of ROBDDs $R_1(x_i), R_2(x_j)$ is nonempty then there exists only if one sequences of 5 neighbor Boolean vectors with the same order of detection of rising and falling transitions in both paths. Obtained results are truth for any vector from a set $\{11,01,11,10,11\}$. If an intersection of ROBDDs $R_1(x_i), R_2(x_j)$ originates vector that contains one don't care on variables x_i, x_j , in this case we have facility to get two Boolean vectors on variables x_i, x_j and, consequently, to get more above mentioned test sequences of 5 neighbor Boolean vectors. If an intersection of ROBDDs $R_1(x_i), R_2(x_j)$ originates vector that contains two don't care on variables x_i, x_j , in this case we have facility to get four Boolean vectors on variables x_i, x_j and,

consequently, still more sequences of 5 neighbor Boolean vectors. The theorem is proved.

Let ROBDD $R_{1s}(x_i), R_{2t}(x_j)$ be correlated with different circuit C outputs that is corresponding paths belong to sub-circuit with outputs s and sub-circuit with output t .

Theorem 3. An intersection of ROBDDs $R_{1s}(x_i), R_{2t}(x_j)$ may be nonempty:

$$R_{1s}(x_i) \& R_{2t}(x_j) \neq \emptyset.$$

Proof. Let consider that a result of an intersection of two cubes, one from each ROBDD from $R_{1s}(x_i), R_{2t}(x_j)$ contains at least one Boolean vector, for example v_1 (on n variables), that belongs to test pair (v_1, v_2) represented by ROBDD $R_{1s}(x_i)$. This test pair is originated by the cube presented by the path from $R_{1s}(x_i)$ root till its 1 terminal node. Vector v_1 at the same time belongs to the cube from $R_{2t}(x_j)$ that is originated by the path from the root $R_{2t}(x_j)$ till its 1 terminal node. This vector is at the same time a test pattern of the test pair for the path corresponding to $R_{2t}(x_j)$. It means that using test fragments originated by vector v_1 we determine delays of path from sub-circuit with output s and sub-circuit with output t . The theorem is proved.

Corollary 1. An intersection of ROBDDs $R_{1s}(x_i), R_{2t}(x_i)$ may be nonempty:

$$R_{1s}(x_i) \& R_{2t}(x_i) \neq \emptyset.$$

In this case test fragment for two paths is derived by matching the same triples. One triple is used for detecting delay of the path corresponding to output s , another – to output t .

Corollary 2. If r ROBDDs have nonempty intersection then there exists if only one test sequence of the length $2r+1$ consisting of neighbor Boolean vectors that detects robust testable PDFs for rising and falling transitions of the corresponding r paths of circuit C . This sequence is characterized by minimal power consumption.

IV. ALGORITHM OF DROWING A CUBE FROM ROBDD AS MUCH AS POSSIBLE CLOSE TO GIVEN CUBE

This algorithm is oriented to cutting Hamming distance between the given cube and a cube drawing from the ROBDD and increasing the number of don't care values in a drawing cube.

Represent the given cube by ROBDD and choose some ROBDD from a set $R_{rob1}, \dots, R_{robL}$. Denote ROBDD representing the given cube as R_1 and ROBDD from this set as R_2 . We try to find the proper cube from ROBDD R_2 .

Edges of ROBDD marked by the same constant 0(1) call edges of the same name, edges marked by the different constants 0,1 call edges of the different names.

Edges running to 0 terminal node we call prohibit ones, other edges call permissible ones.

Edges running to 1 terminal node we call end ones and internal nodes from that of which these edge run call as end nodes.

Note that one of edges running from a node of R_1 is always prohibit one.

We are moving along paths of ROBDD R_2 cutting this procedure because of using ROBDD R_1 . We reach the first

internal node of R_2 (moving at the same time along the only path of R_1) that marked by the same variable that internal node in R_1 . Note these nodes as w_2, w_1 , correspondingly. We form the products k_2, k_1 (cubes) represented by these paths. We correct k_2 adding literals of k_1 that are absent in k_2 and marking the number μ of inverse literals in k_1, k_2 .

Further we continue the similar moving along the path of R_2 to the next internal node w_2 marked by the same variable that node of R_1 and so on till in one of ROBDDs we reach end edge. In this case we come up to 1 terminal node in another ROBDD. If it is R_2 the shortest way is preferable. After that we form products k_1, k_2 , correct k_2 in above mentioned way and store k_2 and μ . Next, we return in the closest branch point of ROBDD R_2 . When a traverse of R_2 was over or we exhausted the given resource, we choose the products k_2 with the lowest μ and among them the product with the lowest rank. Denote the result as k_2^* . It is result of the algorithm as a whole.

Illustrate this procedure by an example. We have circuit C (Fig. 1). Its input variables are ordered: e, b, a, c, d . On Fig. 2 we see ROBDD representing all neighbor test pairs for path $e,5,9$ and on Fig. 3 – ROBDD representing all neighbor test pairs for path $a,1,4,6,8,9$. We choose the shortest path from ROBDD of Fig. 2 and get product $b\bar{a}$. This product forms ROBDD R_1 (Fig. 4). ROBDD of Fig. 3 is R_2 . Analyzing R_2 we try to find product as much as possible close to $b\bar{a}$ by Hamming distance.

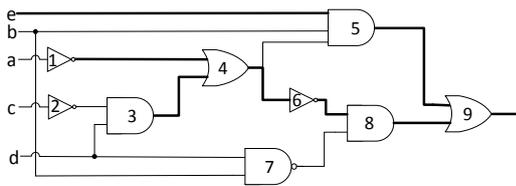


Fig. 1. Circuit C

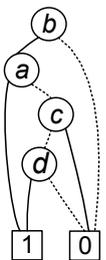


Fig. 2. Path $e,5,9$

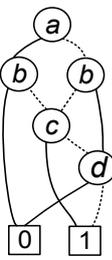


Fig. 3. Path $a,1,4,6,8,9$ (ROBDD R_2)

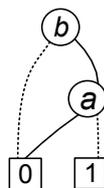


Fig. 4. ROBDD R_1

Thus we move in R_2 from the root along 1 edge to node w_2 marked by variable b . Node w_1 is also marked by variable b . We have: $k_1 = \emptyset, k_2 = e$. The next internal node of R_1 is end one marked by variable a . Then we move from w_2 to 1 terminal node in R_2 forming products: $k_1 = b\bar{a}, k_2 = e\bar{b}c$, corrected k_2 is $e\bar{b}\bar{a}c, \mu = 1$. Choose the closest branch point. In this case, it is root. Now we are moving along 0 edge and reach another node w_2 marked by b . We get $k_1 = \emptyset, k_2 = \bar{e}$. The next internal node of R_1 is end one marked by variable a . Then we move from w_2 along 0 edge to 1 terminal node of R_2 forming products: $k_1 = b\bar{a}, k_2 = \bar{e}\bar{b}c$, corrected k_2 is $\bar{e}\bar{b}\bar{a}c, \mu = 1$. Choose the closest branch point in R_2 : it is the node w_2 marked by b . The next internal node of w_1 in R_1 is end one. Then we move from w_2 along 1 edge to 1 terminal node of R_2

and form products: $k_1 = b\bar{a}, k_2 = \bar{e}\bar{b}\bar{d}$, corrected k_2 is $\bar{e}\bar{b}\bar{a}\bar{d}, \mu = 0$. We got the best Hamming distance and stop traversing, $k_1 \& k_2 = \bar{e}\bar{b}\bar{a}\bar{d}$ (the corresponding cube is 010-0). Consider Boolean vector 01000 from this cube. It turns circuit C into 0, consequently, this vector is b_p test pattern for both paths of circuit C . Note that a_p test pattern for path $e,5,9$ is Boolean vector 11000 and a_p test pattern for path $a,1,4,6,8,9$ is Boolean vector 01100. As a result, we have test sequence of neighbor Boolean vectors:

01000
11000
01000
01100
01000

that detects robust testable PDFs of both rising and falling transitions. In our case first falling then rising transitions for both paths.

V. GENERAL PRINCIPLES OF DERIVING TEST SEQUENCES AND EXPERIMENTAL RESULTS

We have a set of ROBDDs. It is possible different strategies of intersections of ROBDDs. As a result the given set of ROBDDs is separated into sub-sets. If a sub-set contains two or more ROBDDs, it means that its ROBDDs generate the nonempty intersection.

Having executed intersections of ROBDDs for each sub-set we get also ROBDDs. From each of them it is necessary to choose the cube that originates the fragment of test sequence. It is necessary to choose the next fragment as much as possible closer by Hamming distance to previous one.

For circuits from ISCAS'89 we chose at least 10 of the longest path for each circuit output. In table I the information about the considered circuits is given. In the second column the circuit names are marked. In the third, fourth and fifth columns the numbers of inputs (in), outputs (out) and gates (gates) of the circuit are given. In the fifth and sixth columns the numbers of total selected (pt) and robust testable (pr) paths of the circuit are represented. The seventh column shows the ratio (r) of robust testable paths to selected paths (in percentage).

TABLE I. INFORMATION ABOUT THE CONSIDERED BENCHMARKS

No.	Benchmark	in	Out	gates	pt	pr	r
1	s298	17	20	119	146	95	65%
2	s344	24	26	160	159	111	70%
3	s400	24	27	162	258	213	83%
4	s444	24	27	181	237	142	60%
5	s641	54	42	379	309	137	44%
6	s820	23	24	289	232	230	99%
7	s953	45	52	395	338	313	93%
8	s1196	32	32	529	334	162	49%
9	s1488	14	25	653	312	291	93%
10	s1494	14	25	647	336	306	91%

Three algorithms were developed. Each of them contains two the same procedures.

1. The cube for the next fragment is found with using algorithm described in section IV. This algorithm is reduced

to choosing the cube by traversing paths along the ROBDD corresponding to the fragment considered.

2. The cube sequence corresponding to the shortest fragments (triples) is changed by the sequence of Boolean vectors. These vectors are derived by the way being oriented to cut Hamming distance between them [8].

Thus, the differences of the algorithms presented below consist in only the way of intersection of ROBDDs representing test pairs for robust testable path delay faults.

Table II contains control data – results for algorithm without any ROBDDs intersection.

TABLE II. EXPERIMENTAL RESULTS FOR ALGORITHM WITHOUT INTERSECTIONS

No.	Length of the test sequence	Total number of switches	Percentage of one-switches	Peak switches
1	273	331	70%	6
2	312	360	74%	3
3	612	755	69%	6
4	401	471	74%	5
5	350	358	91%	3
6	597	779	64%	6
7	896	1309	51%	7
8	467	652	56%	6
9	823	1146	56%	6
10	877	1242	55%	8

Table III contains results for the algorithm that is based on the pairwise intersection of ROBDDs. If two graphs considered don't have an intersection, the search is performed until the intersection is found, or until all the graphs are examined. In this algorithm the graphs are intersected only inside of each one output sub-circuit. Then the resulting test sequence fragments are combined into one. This algorithm allows reducing the length of the test sequence by 7% approximately and the total number of switches by 10% approximately while slightly increasing the peak values of the switches.

TABLE III. EXPERIMENTAL RESULTS FOR ALGORITHM WITH PAIRWISE INTERSECTIONS

No.	Length of the test sequence	Total count of switches	Percentage of one-switches	Peak switches
1	252	290	77%	3
2	289	326	78%	3
3	568	679	75%	7
4	364	427	76%	5
5	344	362	85%	3
6	579	733	70%	7
7	815	1110	61%	9
8	442	589	63%	7
9	798	1057	64%	7
10	843	1141	62%	6

In the algorithm, the data of which is presented in Table IV, intersections are executed first of all with the graphs for the same subcircuit (the same output but different

inputs). The intersection of the graphs are continued as long as possible with graphs of other subcircuits.

TABLE IV. EXPERIMENTAL RESULTS FOR ALGORITHM WITH INTERSECTIONS OF ROBDDs OF PATHS WITH DIFFERENT INPUTS

No.	Length of the test sequence	Total number of switches	Percentage of one-switches	Peak switches	Highest rank of intersection
1	241	277	79%	5	6
2	274	335	71%	4	9
3	539	642	75%	5	6
4	353	415	75%	5	7
5	328	357	84%	4	11
6	579	715	70%	6	9
7	778	1032	64%	7	9
8	436	605	60%	7	6
9	804	1088	60%	6	7
10	846	1140	61%	7	6

The experimental results show that the last proposed algorithm has the highest potential. On the one hand, it provides the best performance in reducing the length (10% approximately) and the total number of switches (10-15%), while the computational complexity for its implementation is rather low ($O(n)$, where n is the number of selected paths) compared to previous algorithm also based on ROBDD intersections (Table III).

CONCLUSION

The method essentially decreasing power consumption during testing of robust testable PDFs is developed. This result is based on using of ROBDDs R_{rob} properties and their intersections. This method gives additional resources to improve results of RAS technology application.

REFERENCES

- [1] P. Lindgren, M. Kerttu, M. Thornton and R. Drechsler "Low power optimization technique for BDD mapped circuits" ASP-DAC 2001, pp. 615-621.
- [2] R.S. Shelar, S.S. Sapatnekar "An efficient algorithm for low power pass transistor logic synthesis" ASP-DAC 2002, pp. 87-92.
- [3] G. Gekas, D. Nikolos, E. Kalligeros, X. Kavousianos, "Power aware test-data compression for scan-based testing", ICECS 2005. 12th IEEE International Conference on, pp. 1-4.
- [4] J.T. Tudu, E. Larsson, V. Singh, V.D. Agrawal, "On Minimization of Peak Power for Scan Circuit during Test", Test Symposium 2009 14th IEEE European, 2009, pp. 25-30.
- [5] Z. Kotasek, J. Skarvada, J. Strmadel, "Reduction of Power Dissipation Through Parallel Optimization of Test Vector and Scan Register Sequences", IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems, 2010, pp. 364-369.
- [6] V. Sinduja, S. Raghav, J. P Anita, "Efficient don't-care filling method to achieve reduction in test power", ICACCI, 2015, pp. 478-482.
- [7] A.Yu. Matrosova, V.V. Andreeva, E.A. Nikolaeva. Finding Test Pairs for PDFs in Logic Circuits Based on Using Operations on ROBDDs //Russian Physics Journal. 2018. Vol. 61, № 5. pp. 994-999
- [8] A.Yu. Matrosova, V.V. Andreeva, E.A., Tychinskiy V.Z., Goshin G.G. Applying ROBDDs for delay testing of logical circuits. Izvestia vyzov. Physics. 2019. v. 62, № 5. pp. 86-94.
- [9] A. Matrosova, V. Lipsky, A. Melnikov, V. Singh. Path delay faults and ENF. IEEE East-West Design & Test Symposium. St. Petersburg: IEEE, 2010, pp. 164-167.

Development of Modified Block Cipher Algorithm TEA, Free from Vulnerability of “Connected Keys Attack”

Sergey A. Klyokta

*Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
seregaklyokta@mail.ru*

Alexander I. Zhukov

*Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
zhukov000@gmail.ru*

Andrey G. Lobodenko

*Information Systems and Radioengineering
Don State Technical University
Rostov-on-Don, Russia
andrey@sssu.ru*

Nikita I. Chesnokov

*Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
4esnog96@gmail.com*

Larissa V. Cherckesova

*Mathematics and Computer Sciences Chair
Don State Technical University
Rostov-on-Don, Russia
chia2002@inbox.ru
ORCID 0000-0002-9392-3140*

Irina A. Pilipenko

*Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
ipilipenko@donstu.ru*

Olga A. Safaryan

*Information System Cybersecurity Chair
Don State Technical University
Rostov-on-Don, Russia
safari_2006@mail.ru*

Vitaliy M. Porksheyayn

*Applied Mathematics Chair
Don State Technical University
Rostov-on-Don, Russia
spu-46@donstu.ru*

Abstract— *In the framework of this article, general information on the block cipher TEA was reviewed, including the history of its creation, advantages and disadvantages, as well as details of the process of encryption and decryption. Three main vulnerabilities of the cipher are described: differential cryptanalysis, the presence of equivalent keys, and in most detail – the "related-key" attack. Next, we consider the algorithm for generating the DES round keys and its modification, which makes it possible to integrate this algorithm into the TEA.*

Keywords: *cryptoalgorithm, block cipher, keys of ciphering, cryptanalysis, programming language, Python.*

I. INTRODUCTION

The problem of information protecting by transforming it, excluding its reading by an unauthorized person, worried the human mind from a long time ago.

Why the problem of using cryptographic methods in information systems has become particularly relevant at the moment? Everything is simple – this is due to the fact that information technology every day more and more pervades all sides of our lives, due to which we are forced to store information of different levels of confidentiality and exchange it.

Therefore, there is an urgent need to ensure the security of transmitted and stored data. It is for this purpose that various types of information encryption are widely used in these systems. Encryption is a reversible transformation of information for the purpose of hiding from unauthorized persons, while at the same time providing authorized users with full access to it. Currently, there are a huge number of algorithms that encrypt information to prevent its falling into the hands of those who do not have the right to access it. Among these ciphers is a large layer of algorithms based on Feistel networks.

The Feistel network is one of the methods for building block ciphers, consisting of cells called Feistel cells. The input of each cell receives data and a key. At the output of each cell, the modified data and the modified key are received. All cells are the same type and say that the network is a certain repetitive structure. The key is selected

depending on the encryption / decryption algorithm and changes when you move from one cell to another.

During encryption and decryption, the same operations are performed, only the order of the keys differs. In view of the simplicity of operations, the Feistel network can be easily implemented both programmatically and in hardware.

One of these ciphers is the block cipher TEA, which is one of the simplest in its family.

II. FORMULATION OF THE PROBLEM

The purpose is to develop a modification of the block cipher TEA, which will be devoid of vulnerability to attacks on "related keys". To achieve this goal, a number of tasks were identified:

- analyze the original algorithm and its existing modifications;
- explore the general principles of attacks on related keys, as well as their implementation in relation to the TEA algorithm;
- modify the algorithm using known principles of information systems protection from cyberattacks.

III. THEORETICAL FOUNDATION

Tiny Encryption Algorithm (TEA) [1] is block encryption algorithm of "Feistel Network" type. This algorithm was developed at Cambridge University Computer Science Department by David Wheeler and Roger Needham. It was first introduced in 1994 at Symposium on Fast Encryption Algorithms in the Belgian city of Leuven [2].

The cipher is not patented, it is widely used in many cryptographic applications and a wide range of hardware. A considerable level of its demand is due to extremely low requirements for memory and ease of implementation. The algorithm has both a software implementation in different programming languages and a hardware implementation on integrated circuits such as FPGA.

The TEA encryption algorithm [3] is based on bit operations with a 64-bit block has a 128-bit encryption key. For this cipher, the standard number of rounds is 64,

which corresponds to 32 cycles, but in order to obtain the best encryption or performance increase, the number of cycles can vary from 8 (16 rounds) to 64 (128 rounds). As an operation of superposition, modulo 2^{32} additions is used – this is due to the asymmetry of the Feistel Network.

Advantages of Cipher. The positive characteristics of the TEA block cipher are its ease of implementation, small code size and rather high execution speed, as well as the possibility of optimizing execution on standard 32-bit processors, since the XOR, bit-shift, and addition modulo 2^{32} . Since the algorithm does not use substitution tables and the round-robin function is fairly simple, the algorithm requires at least 16 cycles (32 rounds) to achieve efficient diffusion, although it is full I diffusion is achieved over the 6 cycles (12 rounds) [3]. The algorithm has excellent resistance to linear cryptanalysis and is quite good to differential.

Description of the algorithm. The source text is divided into blocks of 64 bits each. The 128-bit K key is divided into four 32-bit sub keys K_0 , K_1 , K_2 , and K_3 . This completes the preparatory process, after which each 64-bit block is encrypted for 32 cycles (64 rounds) according to the algorithm given [4].

Suppose that the right and left sides (L_n , R_n) enter the input of the n th round (for $1 \leq n \leq 64$), then the left and right sides (L_{n+1} , R_{n+1}) at the output of the n -th round, which are calculated according to the following rules:

$$L_{n+1} = R_n.$$

If $n = 2 * i - 1$ for $1 \leq i \leq 32$ (odd rounds), then

$$R_{n+1} = L_n \boxplus (\{[R_n \ll 4] \boxplus K_0\} \oplus \{R_n \boxplus i * \delta\} \oplus \{[R_n \gg 5] \boxplus K_1\}).$$

If $n = 2 * i$ for $1 \leq i \leq 32$ (even rounds), then

$$R_{n+1} = L_n \boxplus (\{[R_n \ll 4] \boxplus K_2\} \oplus \{R_n \boxplus i * \delta\} \oplus \{[R_n \gg 5] \boxplus K_3\}),$$

where $X \boxplus Y$ — Addition of numbers X and Y by module 2^{32} .
 $X \oplus Y$ — Bitwise exclusive "OR" (XOR) of the numbers X and Y , which in the C programming language is denoted by $X \wedge Y$.

$X \ll Y$ and $X \gg Y$ — Bitwise bit shift operations X to Y bits left and right respectively.

The constant δ was deduced from the Golden section

$$\delta = (\sqrt{5} - 1) * 2^{31} = 2654435769_{10} = 9E3779B9_h [5].$$

In order to prevent attacks based on the symmetry of rounds, in each round, constant is multiplied by number of the cycle i .

It is also obvious that in the TEA encryption algorithm there is no algorithm for scheduling keys as such. Instead, in odd rounds, we use K_0 and K_1 subkeys; in even rounds we use K_2 and K_3 .

Crypto analysis. It is assumed that this algorithm provides a security comparable to the IDEA encryption algorithm, since it uses the same idea of using operations from orthogonal algebraic groups [6]. This approach perfectly protects against the methods of linear cryptanalysis.

IV. ATTACKS ON RELATED KEYS

The algorithm is most vulnerable to "attacks on related keys", due to a simple key schedule (including the lack of an algorithm for scheduling keys as such).

What are these attacks on related keys?

A key – based attack is a type of cryptographic attack in which a cryptanalyst can observe the operation of an encryption or decryption algorithm that uses several

secret keys. This attack is not a simple search for all possible key values. Initially, the cryptanalyst does not know anything about the exact meaning of the keys, but it is assumed that the attacker knows some mathematical relation connecting the keys to each other. For example, the relationship can be simply a value of XOR with a known constant or a more complex functional relationship. In real life, such dependencies can occur when there is a failure in the hardware or poorly designed security protocols.

Differential cryptanalysis. TEA is quite resistant to differential cryptanalysis. Attacking 10 rounds of TEA requires $2^{52.5}$ selected open texts and has a time complexity of 2^{84} [7].

The best result is a cryptanalysis of 17 rounds of TEA [8]. This attack requires only 1920 selected open texts, but has a temporary complexity $2^{123.37}$.

Equivalent keys. Another weakness of the TEA algorithm is the presence of equivalent keys. It was found that each key has three equivalent [9] sub keys.

This means that effective key length is only 126 bits instead of 128 designed by the developers. For this reason, TEA should not be used as a hash function, as reflected in Andrew Huang's "Hacking the Xbox: an introduction to reverse engineering" book by hacking the Microsoft Xbox game console.

TABLE I. EQUIVALENT KEYS

K_0	K_1	K_2	K_3
K_0	K_1	$K_2 + 80000000_h$	$K_3 + 80000000_h$
$K_0 + 80000000_h$	$K_1 + 80000000_h$	K_2	K_3
$K_0 + 80000000_h$	$K_1 + 80000000_h$	$K_2 + 80000000_h$	$K_3 + 80000000_h$

Since the TEM is a block cipher algorithm, in which the block length is 64-bits, and the data length can be not a multiple of 64-bits, the values of all bytes complementing the block to a multiplicity of 64 bits are set to 0×01 .

Consideration of the selected vulnerability. There are several known attacks on TEA, which exploit the vulnerability of related keys. To understand the essence of this vulnerability, consider the simplest attack [10].

Let there be a key $K = (K_0, K_1, K_2, K_3)$. Change the keys K_2 and K_3 bits at position 30 (following, after the most significant bits). With a probability of approximately 0.5, the output of even Feistel network rounds that use the changed keys will match the output on the original keys. This fact already gives a two-round cyclic differential characteristic with a probability of 0.5, and accordingly a 60-round characteristic with a probability of 2^{-30} .

Thus, for hacking a 64-round TEA algorithm, one pair of "matched" keys and 2^{34} plaintexts are enough. The possibility of such an attack is due to the fact, that the TEA algorithm does not involve any development of round keys, but simply uses 4 sub keys throughout all rounds.

V. PRODUCED MODIFICATIONS

The obvious solution to the problem described above is to add an algorithm for generating round keys. Keys should not be created identically, linearly. In the ideal case, each bit of the key should affect all rounds and each round in different ways [11].

In general, the requirements for good round – key generation algorithms strongly overlap with requirements for cryptographic hash functions.

First, the stages (iteration) of the algorithm must be difficult to invertible. That is, having a key of any one round, it should be difficult to get information about the bits of the key of any other round.

Secondly, in order to avoid the formation of identical keys for different rounds, the key generation algorithm must be protected from collisions.

Finally, it is not possible to produce controlled changes to the created rounds. It was controlled changes in the keys that allowed us to obtain a differential characteristic in the attack described earlier.

Also, in the keys there should be no so-called "dead zones". Each bit of the key should almost equally affect the transformations produced by the whole key. In other words, there should not be any bits in the round key that do not actually participate in the encryption process.

It is well known, that DES's algorithm has a good schedule of keys [12]. Although the set of round keys is the result of simple linear transformations, these keys are fairly reliable and resistant to attacks on related keys. Provided, of course that the encryption key does not choose one of the known weak or partially weak keys [13].

This combination of simplicity and reliability is exactly what the TEA cipher possesses in general, and what is so lacking in its round keys. It is because of these qualities that it was decided to modify the original TEA by adding an adapted version of the algorithm for generating round keys used in DES. The original DES key encryption algorithm includes several steps (Figure 1):

1. An input key of 56 bits (7 bytes) is input. It is complemented by the check bits at positions 8, 16, 24, 32, 40, 48, 56, 64 so that each byte has an odd number of units.

```

74 def prepare_initial_password(password: str) -> Tuple[bytes, int]:
75     if type(password) == str:
76         password = password.encode('utf-8')
77         septs = bytes_to_septs(password[:7])
78         res = []
79         parity_bits = 0
80         for sept in septs:
81             curr_byte, new_bit = append_parity_bit(sept)
82             res.append(curr_byte)
83             parity_bits = (parity_bits << 1) | new_bit
84         return bytes(res).ljust(8, b'\x00'), parity_bits
85
86
87

```

Fig. 1. Function of expanding the source key with parity bits.

2. An initial permutation is applied to extended key (length 64), given by a table of 56 elements, so that in the sequence that passed the permutation, the bit whose number corresponds to the cell number of the table is in the initial (extended) key a bit with the number specified in the current cell (Figure 2).

```

6 # Initial permut made on the key
7 CP_1 = [57, 49, 41, 33, 25, 17, 9,
8         1, 58, 50, 42, 34, 26, 18,
9         10, 2, 59, 51, 43, 35, 27,
10        19, 11, 3, 60, 52, 44, 36,
11        63, 55, 47, 39, 31, 23, 15,
12        7, 62, 54, 46, 38, 30, 22,
13        14, 6, 61, 53, 45, 37, 29,
14        21, 13, 5, 28, 20, 12, 4]
15
16

```

Fig. 2. Initial permutation.

As you can see, in addition to mixing bits, during the permutation, the key length also decreases. This permutation discards the verification bits mentioned in:

1. This table is known in advance and is used by the cipher regardless of the key or plain text.

2. Further, within 16 cycles, directly round keys are developed. At each iteration, the key is represented as two

blocks of 28 bits each. Each of the blocks is cyclically shifted to the right by 1 or 2 bits according to the shift table.

3. Next, the combined blocks undergo a second permutation, during which 56 bits are selected 48, which constitute the key of the round. Accordingly, 16 keys of length 48 are generated during the sixteen iterations (Figure 3).

```

17 # Permut applied on shifted key to get Ki+1
18 CP_2 = [14, 17, 11, 24, 1, 5, 3, 28,
19         15, 6, 21, 10, 23, 19, 12, 4,
20         26, 8, 16, 7, 27, 20, 13, 2,
21         41, 52, 31, 37, 47, 55, 30, 40,
22         51, 45, 33, 48, 44, 49, 39, 56,
23         34, 53, 46, 42, 50, 36, 29, 32]
24
25

```

Fig. 3. Second permutation applied at each iteration of key generating.

From the description of the original algorithm it is clear that although it is quite simple, in our case it still needs to be modified. The TEA algorithm assumes the use of round keys of length 32, rather than 48. In addition, 16 keys are too small: we need to be able to create from 16 to 128 keys. To satisfy these requirements, the original DES algorithm has been improved. But, since it is reliable in the original form, it was decided not to change it, but to supplement:

1. Find a way to increase the length of the round key from 48 to 64. This will, in fact, create two keys of 32 bits per iteration of the round key algorithm.

2. Increasing the length of the key to be produced, only through the additional 16 bits, make the key variance, which would allow creating up to 64 pairs of unique round keys (since in one iteration of the algorithm two keys are produced at once, in total we get 128 keys we need) [14].

The original algorithm of the key schedule describes 8 check bits that do not participate in encryption in any way (Figure 4).

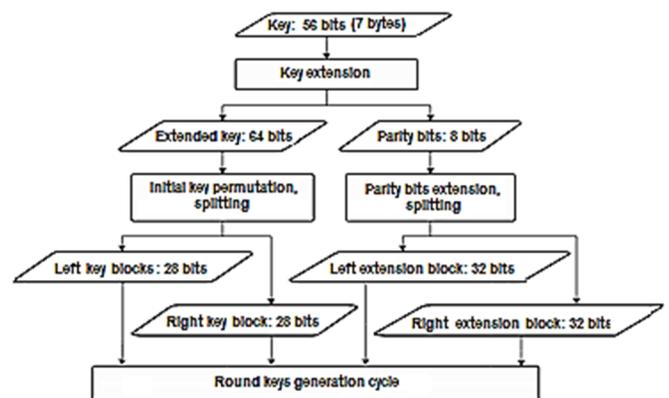


Fig. 4. The initial stage of the keys generation algorithm.

To expand the size and number of final keys in a modified algorithm, it was decided to use them (Figure 5).

```

112 def generatekeys(password: str, cycles=64) -> List[Tuple[int, int]]:
113     keys = []
114     init_password, parity_bits = prepare_initial_password(password)
115
116     # Parity Bits Array
117     pba = string_to_bit_array([parity_bits])
118     pba = permut(pba, PB_EXTENSION)
119     pba_l, pba_r = nsplit(pba, 32)
120
121     # Original DES key
122     key = string_to_bit_array(init_password)
123     key = permut(key, CP_1)
124     key_l, key_r = nsplit(key, 28)
125
126

```

Fig. 5. Implementation of algorithm initial stage for keys generating.

At the initial permutation stage, before the key generation cycle, an extension block is created. To do this, using a spreadsheet that we selected, we obtain a length block of eight (8) test bits. Further, this block is divided into two equal sub blocks (left and right). Initial transformations of the original algorithm remain unchanged at this stage (Figures 6 and 7).

```

36 # Matrix to extend parity byte
37 PB_EXTENSION = [
38   0, 1, 2, 3, 4, 5, 6, 7,
39   5, 6, 0, 2, 7, 3, 4, 1,
40   1, 0, 4, 3, 6, 7, 2, 5,
41   6, 4, 7, 5, 1, 0, 3, 2,
42   7, 5, 3, 6, 2, 1, 0, 4,
43   5, 2, 7, 6, 3, 4, 1, 0,
44   2, 3, 0, 1, 7, 4, 5, 6,
45   4, 7, 1, 0, 2, 6, 3, 5,
46 ]
47 ]

```

Fig. 6. Spreading bit expansion table.

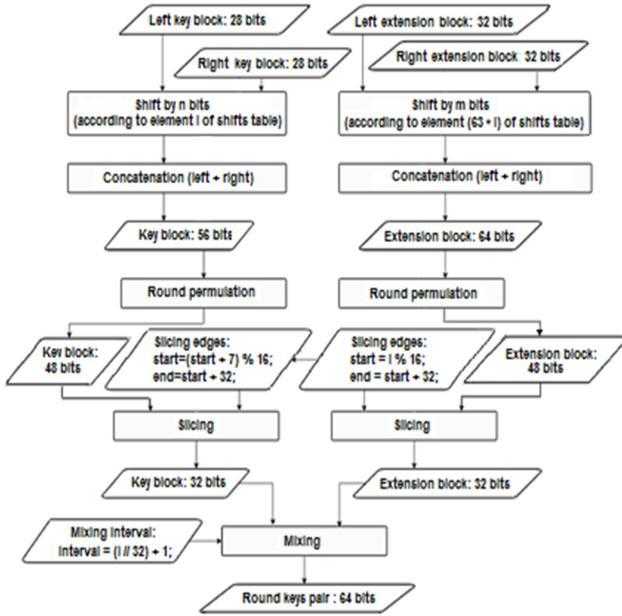


Fig. 7. Cycle of generation of round keys.

At each i -th iteration of the key generation, both the key blocks (as in the original algorithm) and the extension blocks, are shifted, according to the shift table. However, to exclude the generation of identical keys, the shift table itself was increased 4 times. The first 16 shifts remained unchanged, and the subsequent segments of 16 shifts are obtained by multiplying each element by $(i \text{ div } 16) + 1$, and then mixing some additional elements and integer segments.

The shift of the key block determines the element i of the shift table, and the shift of the expansion block is the element $(63 - i)$ (Figures 8 and 9).

```

126 for i in range(cycles):
127     key_l, key_r = shift_pair(key_l, key_r, SHIFT[i])
128     pba_l, pba_r = shift_pair(pba_l, pba_r, SHIFT[63 - i])
129
130     tmp_key = key_l + key_r
131     tmp_pba = pba_l + pba_r
132
133     pba_start = i % 16
134     key_start = (pba_start + 7) % 16
135
136     key_part = permute(tmp_key, CP_2)[key_start:key_start + 32]
137     pba_part = permute(tmp_pba, CP_2)[pba_start:pba_start + 32]
138
139     key_pair = mix(key_part, pba_part, (i // 32) + 1)
140     keys.append((bits_arr_to_int(key_pair[:32]), bits_arr_to_int(key_pair[32:64])))
141
142 return keys
143 ]

```

Fig. 8. Implementing the round-robin generation cycle.

```

26
27 # Matrix that determines the shift for each round of keys
28 SHIFT = [
29   1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 1,
30   # Extension
31   4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 3, 3, 4, 4, 3, 4,
32   3, 2, 3, 3, 2, 2, 3, 3, 3, 3, 3, 3, 2, 3, 3,
33   4, 5, 5, 4, 5, 5, 5, 5, 5, 5, 5, 4, 5, 5, 4,
34 ]
35 ]

```

Fig. 9. Extended shift table.

1. The shifted blocks are concatenated in pairs (left and right).
2. Blocks are subjected to a round permutation in the same

way as in the original algorithm.

3. Select the boundaries of trimming blocks. For the expansion block, the beginning and end of the cropping are given by the following formulas:

$$H_p = i \text{ mod } 16, K_p = H_p + 32.$$

For the key block:

$$H_k = (H_p + 7) \text{ mod } 16; K_k = H_k + 32,$$

where H_p is the beginning of the trimming of the expansion block, K_p is the end of the trimming of the expansion block, H_k is the start of the trimming of the key block, K_k is the end of the trimming of the key block.

4. Blocks are cut according to the selected boundaries. At this stage, both blocks have a length of 32.
5. Select the mixing interval: $I = (i \text{ div } 32) + 1$ (Figure 10).

```

102
103 def mix(arr1: List[Any], arr2: List[Any], period=1) -> List[Any]:
104     result = []
105     sum_length = len(arr1) + len(arr2)
106     for i in range(sum_length // period):
107         l_border = i * period
108         r_border = l_border + period
109         result += arr1[l_border:r_border] + arr2[l_border:r_border]
110     return result
111 ]

```

Fig. 10. Block Blending Function.

The key block with the expansion unit is combined (and mixed) according to the selected expansion interval.

The meaning of interval is how the elements of both blocks are mixed. For example, if the key block is $K = k_i$, extension block is $E = e_i$, then for $i = 1$, elements will be intermixed as follows:

$$\{k_0, e_0, k_1, e_1, \dots\};$$

$$\text{for } i = 2: \{k_0, k_1, e_0, e_1, \dots\}.$$

The combined and mixed block has a length of 64. After dividing it in half, we get a pair of ready-made round keys.

VI. RESULTS AND CONCLUSIONS

To measure reliability and efficiency of developed algorithm, several simple program tests were developed and used [15].

The first test measures the time for full encryption and decryption (including the generation of keys) on data of different sizes (from 10Kb to 1Mb). The code and test results are shown in the following Figures 11 and 12.

```

15
16 def timing_test(init_plain_data: str, key: str):
17     init_size_kb = 10
18     plain_text_sizes = [1, 2.5, 5, 7.5, 10, 25, 50, 75, 100]
19     init_len = len(init_plain_data)
20     print('Timings.\n')
21     for koef in plain_text_sizes:
22         plain_data = repeat(init_plain_data, math.floor(init_len * koef))
23
24         start_enc = t()
25         encrypted_data = encrypt(plain_data, key)
26         start_dec = t()
27         decrypt(encrypted_data, key).decode('utf-8')
28         end = t()
29
30         enc_time = (start_dec - start_enc) * 1000
31         dec_time = (end - start_dec) * 1000
32
33         kb_size = init_size_kb * koef
34         print('\n%dkb. Encrypt: %fms; Decrypt: %fms.' % (kb_size, enc_time, dec_time))
35

```

Fig. 11. Code time test for full encryption and decryption.

```

nikita@nikita-mint ~/dev/learn/tea $ python3 main.py -p resources/text10kb.txt -k the_key -i --run-tests
Timings.
10kb. Encrypt: 89.118491ms; Decrypt: 88.634046ms.
25kb. Encrypt: 217.409021ms; Decrypt: 217.957396ms.
50kb. Encrypt: 423.496665ms; Decrypt: 433.376475ms.
75kb. Encrypt: 636.746577ms; Decrypt: 652.200219ms.
100kb. Encrypt: 857.613287ms; Decrypt: 862.340708ms.
250kb. Encrypt: 2120.288370ms; Decrypt: 2167.841541ms.
500kb. Encrypt: 4237.158778ms; Decrypt: 4317.270937ms.
750kb. Encrypt: 6344.596065ms; Decrypt: 6463.521644ms.
1000kb. Encrypt: 8488.870506ms; Decrypt: 8697.158183ms.

```

Fig. 12. Output of the full time measurement of encryption and decryption

As you can see on graphs below, with increase in amount of data, the running time of the algorithm grows linearly. At minimum, this indicates that the modified algorithm for time indicators is not inferior to the original TEA (Figures 13 and 14).

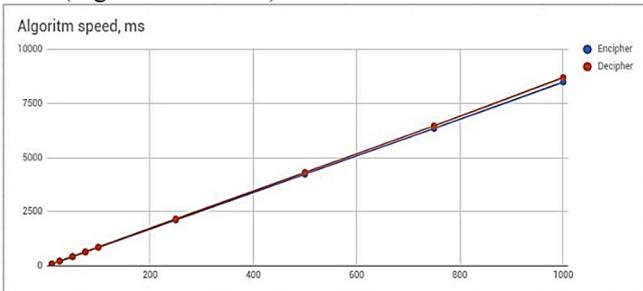


Fig. 13. Graph of changes in the speed of encryption and decryption.

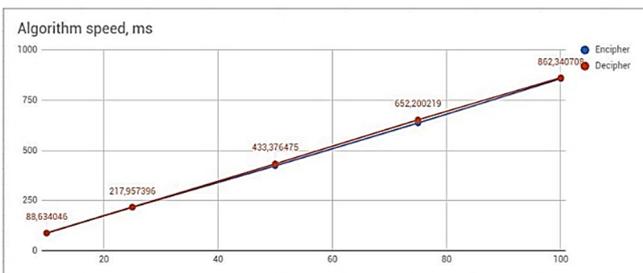


Fig. 14. Graph of the change in the speed of encryption and decryption, depending on the amount of data (from 10Kb to 100Kb).

Further, the key generation algorithm was tested separately. As you can see in the following figure, the production of 10,000 sets of round keys from the original key "the key" in total took 24062 ms. In other words; the generation of one set on average takes 2.4 ms (Figures 15–17).

```

66
67 def test_genkeys(pwd: str, times=10000):
68     start = t()
69     for i in range(times):
70         generatekeys(pwd)
71     full_time = (t() - start) * 1000
72     print('Keys generation (%d times) timing: %fms.' % (times, full_time))
73

```

Fig. 15. Test of the algorithm for generating round keys.

```

>>> test_genkeys('the_key')
Keys generation (10000 times) timing: 24062.054739ms

```

Fig. 16. Testing the algorithm for generating round keys

To check the security of algorithm against attacks related to the presence of equivalent keys, the following test was used.

```

40
41 def equal_keys_test(key: str) -> List[Tuple[str, int, str, int]]:
42     keys = _flatten_keys(generatekeys(key))
43     keys_count = len(keys)
44     equals = []
45     print('Start equal keys test...')
46     for i in range(keys_count):
47         left_key = keys[i]
48         left_eq_key = left_key ^ 0x80000000
49         for j in range(i + 1, keys_count):
50             right_key = keys[j]
51             right_eq_key = right_key ^ 0x80000000
52             ti = (
53                 left_key == right_key or
54                 left_key == right_eq_key or
55                 left_eq_key == right_key or
56                 left_eq_key == right_eq_key
57             )
58             left_num = '%d:%d' % (i // 2, i % 2)
59             right_num = '%d:%d' % (j // 2, j % 2)
60             equals.append((left_num, left_key, right_num, right_key))
61
62     if len(equals) > 0:
63         print('Found several equal keys:\n')
64         for entry in equals:
65             print('%s - %d; %s - %d' % entry)
66     else:
67         print('No equal keys found.')
68     return equals
69

```

Fig. 17. Test of the presence of pairs of equivalent keys among all the round keys from one selection.

Even for the most basic and unsafe encryption keys, algorithm generates 128 unique, nonequivalent keys, as shown in Figure 18.

```

nikita@nikita-mint ~/dev/learn/tea $ python3 main.py -m encrypt --run-tests -d -k aaaaaa -i
No equal keys found.
nikita@nikita-mint ~/dev/learn/tea $ python3 main.py -m encrypt --run-tests -d -k 111111 -i
No equal keys found.

```

Fig. 18. Test results for the presence of pairs of equivalent keys.

Let us analyze this result. The vulnerability of the original TEA is due to the fact that when the older bits of some round keys were changed, the encryption result did not change. In our algorithm, changing any bit of the source key leads to a change in the set of parity bits, which in turn leads to a significant change in the entire set of round keys. In other words, in its pure form, the old vulnerability can already be considered closed.

Thus, with minimal time and resource losses, we obtained a more reliable modification of the TEA algorithm.

REFERENCES

- 1] http://citforum.ru/internet/infsecure/its2000_19.shtml
- 2] S. Panasenko, "Encryption algorithms. Special reference book". – St. Petersburg: BHV – Petersburg, 2009, 576 Pp.
- 3] J. Kelsey. "Key-schedule cryptanalysis of IDEA, G-DES, GOST, SAFER, and Triple-DES". / B. Schneier, D. Wagner, "Lecture Notes in Computer Science" 1109: magazine. – 1996 – Pp. 237–251. DOI: 10.1007 / 3-540-68697-5_19
- 4] M. Dunkelman, J. Orr, "Related-Key Attacks". / Department of Computer Science, University of Haifa, Faculty of Mathematics and Computer Science, Weizmann Institute of Science: Journal. 2011 – No.1. Pp. 1–10.
- 5] A.V. Babash, G.P. Shankin, "Cryptography". – M.: SOLON-PRESS, 2007. – 512 p. (In Russian).
- 6] Kelsey, J. Related-key cryptanalysis of 3-WAY, Biham-DES, CAST, DES-X New DES, RC2, and TEA / Schneier, B., Wagner David., Lecture Notes in Computer Science 1334: Journal. – 1997. Pp. 233–246. DOI: 10.1007 / Bf0028479.
- 7] E. Biham New types of cryptanalytic attacks using related keys // Journal. Springer-Verlag: 1994. No. 4. Pp. 229–246. DOI: 10.1007 / Bf00203965.
- 8] Free encyclopedia [Electronic resource]. – Access mode: <https://ru.wikipedia.org/wiki>

- [9] D. Moon, "Impossible differential cryptanalysis of reduced round XTEA and TEA" / K. Hwang, W. Lee, S. Lee, J. Lim, "Lecture Notes in Computer Science". 2365: Journal. — 2002. — Pp. 49–60. DOI: 10.1007/3-540-45661-9_4
- [10] J. Sammons, M. Cross, "The Basics of Cyber Safety. Computer and Mobile Device Safety Made Easy". 1-st Edition Syngress. 2016.–254 p.
- [11] A. Skavhaug, J. Guiochet, E. Schoitsch, F. Bitsch, "Computer Safety, Reliability and Security". SAFECOMP.2016. Workshops. 2016.–400 p.
- [12] M.S. Kosyakov, "Introduction to Distributed Computing". St. Petersburg, 2014. – 155 p. (In Russian).
- [13] A. Boyle, and N. Panko, "Corporate Computer Security", III-th Edition. Prentice Hall. 2013. – 661 p.
- [14] J. Stallings and M. Brown, "Computer Security: Principles and Practice", III-th Edition. Prentice Hall. 2014. – 820 p.
- [15] N.M. Deytel, "Operating systems". V.2. "The distributed systems, networks, safety" / N.M. Deytel, P.D. Deytel, D.R. Hofnes; M.: BINOMIAL, 2013. – 704 p. (In Russian).

Masking Internal Node Faults and Trojan Circuits in Logical Circuits

A. Matrosova
Tomsk State University
Tomsk, Russia
mau11@yandex.ru

V. Provkin
Tomsk State University
Tomsk, Russia
prowkan@mail.ru

E. Nikolaeva
Tomsk State University
Tomsk, Russia
nikolaeve-ea@yandex.ru

Abstract—A combinational circuit C is considered. Masking of internal node logical faults with using the sub-circuit that outputs are connected with circuit C internal nodes that are fed by fault nodes is suggested. The sub-circuit inputs are connected with either circuit C inputs or with internal nodes of circuit C that precede the fault nodes. Masking is based on applying of incompletely specified Boolean functions of internal nodes. Algorithms of deriving incompletely specified Boolean function for some internal node v are described. One of them gets the incompletely specified function that depends on input variables of circuit C , another gets the incompletely specified Boolean function that depends on internal variables of circuit C that corresponds to internal nodes preceding fault nodes. Using these algorithms for several fault nodes we obtain the system of incompletely specified Boolean functions that is implemented by masking (patch) circuit. This approach may be also applied for masking Trojan Circuits (TCs). It is supposed that TC output is injected into a line of combinational circuit C . Experimental results are given. They demonstrate possibilities of essential cutting overhead when using patch function in comparison with duplication.

Keywords—combinational circuits, logical faults, Trojan circuits, ternary simulation, Reduced Ordered Binary Decision Diagram (ROBDD), incompletely specified Boolean functions

I. INTRODUCTION

Masking logical faults of internal nodes of combinational circuit C is suggested. These faults may manifest themselves either on the last stages of circuit C fabrication or under injection of Trojan Circuits (TCs) when TC output is inserted in the combinational circuit C line. It is considered that masking circuit outputs (patch circuit outputs) are connected with circuit C internal nodes that are fed by fault nodes and masking circuit inputs are connected with either circuit C inputs or with internal nodes of circuit C preceding the fault nodes. This assumption corresponds to Engineering Change Order (ECO) technologies that are applied for combinational circuit correction. The last results in the frame of ECO technologies that are close to our investigations are represented in [1, 2]. The results in [1, 2] are based on finding some implementation of system of incompletely specified Boolean functions [3, 4] corresponding to fault nodes. For that they use one of SAT systems. The obtained implementation then is applied to derive masking (patch) circuit. Patch circuit may be obtained with using one of CAD tools. Note that fault nodes rather often originate the system of poorly determined incompletely specified Boolean functions. It means that for CAD tools just such systems are preferable to derive as much as possible simpler patch circuit. That is why our approach is oriented on applying system of incompletely specified Boolean functions. Each function of this system we represent by two ROBDDs, one ROBDD presents on-set another off-set of incompletely specified Boolean function. We first studied facilities of getting incompletely specified Boolean functions for circuits

from MCNC system. It seemed that such functions may be obtained for circuits having several tens and even about two hundreds inputs (comp – 32 inputs, term1 - 34 inputs, C432 - 36 inputs, too_large - 38 inputs, i3 -135 inputs, pair – 173 inputs, C5315 – 178 inputs, i4 – 192 inputs). Certainly, we cannot obtain incompletely specified Boolean functions for nodes of multipliers but for many other circuits it is possible. Preliminary experiments on circuits from MCNC show that even if we use incompletely specified Boolean functions, we may get masking (patch) circuit that complexity is close to duplication of sub-circuit that of which outputs are fault nodes. The closer fault nodes to inputs of circuit C the more probable such situation. It means that it is necessary to compare obtained patch circuits with duplication.

In section II the problem statement is given. In section III the algorithm of getting incompletely specified Boolean function of internal node v depending on input variables of combinational circuit C is described. In section IV mapping above mentioned incompletely specified Boolean function on internal variables of circuit C is described. In Section V a way of deriving patch circuit from system of incompletely specified Boolean functions is suggested. Experimental results are given in Section VI.

II. PROBLEM STATEMENT

We have a combinational circuit and a set of logical fault nodes. It is necessary to get masking circuit for these faults in the frame of Engineering Change Ordering (ECO) technologies. We derive masking circuit connected either with input variables of circuit C or circuit C internal variables corresponding to its internal nodes preceding the fault nodes. In both cases outputs of masking circuits are connected with nodes that are fed by the fault internal nodes. We try to get masking circuit as simple as possible using incompletely specified Boolean functions for fault nodes. Actually, incompletely specified Boolean function for fault node provides the best resurces for getting the simplest masking (patch) circuit with using any concrete CAD tool. When we want to derive masking (patch) circuit that inputs are connected with the internal nodes preceding the fault nodes, we have to map incompletely specified function of node v depending on input variables of circuit C into the internal variables corresponding to these internal nodes. Note that Boolean simulation for mapping is impossible for circuits with several tens and more inputs. In this paper we suggest to get the incompletely specified Boolean functions on internal variables of circuit C using operations on ROBDDs derived from fragments of combinational circuit (combinational part of a sequential circuit) and special ternary simulation suggested by us at the beginning of two thousand years. As we know this way of mapping is suggested for the first time. Facility of this approach are restricted by complexity of fragments of a combinational circuit but the approach allows executing mapping of incompletely

specified Boolean functions (using CUDD) for circuits depending on several tens and more variables.

III. DERIVING INCOMPLETELY SPECIFIED BOOLEAN FUNCTIONS DEPENDING ON INPUT VARIABLES OF CIRCUIT C

Take into consideration that for each internal node v of circuit C we may derive incompletely specified Boolean function f_v . All test patterns for stuck at 0 fault of node v is on-set $M_1(f_v)$ of this function and all test patterns for stuck at 1 fault of node v is off-set $M_0(f_v)$ of this function [5].

Consider some incompletely specified Boolean function f_1 and completely specified Boolean function f_2 , that are represented by their on-sets and off-sets depending on the same variables: $M_1(f_1), M_0(f_1); M_1(f_2), M_0(f_2)$.

Definition. Function f_2 implements function f_1 , if $M_1(f_2)$ contains $M_1(f_1)$ and $M_0(f_2)$ contains $M_0(f_1)$.

For each node v of circuit C we may derive completely specified Boolean function $\varphi(x_1, x_2, \dots, x_n)$ that is implemented by the sub-circuit C_v of circuit C . Sub-circuit C_v inputs are inputs of circuit C and the sub-circuit C_v output is node v . Note that function $\varphi(x_1, x_2, \dots, x_n)$ is implementation of incompletely specified Boolean function f_v . Any implementation of f_v may be applied as masking sub-circuit for fault of node v , but we want to find the better implementation if possible.

For finding incompletely specified Boolean function f_v , we use Reduced Ordered Binary Decision Diagrams (ROBDDs) and operations on them. Remind that these operations are characterized by polynomial complexity.

First we derive ROBDD $R(C_v^i)$ for sub-circuit C_v^i . This sub-circuit corresponds to i -th output of circuit C under condition that node v is the input of the sub-circuit together with variables x_1, x_2, \dots, x_n . ROBDD $R(C_v^i)$ represents function $f_i(v, x_1, \dots, x_n)$ while variable v is used as the first variable under Shannon decomposition. Further we find Boolean difference $D_v f_i$ of this function on variable v , representing $D_v f_i$ by ROBDD. Note that $D_v f_i$ presents at the same time observability of node v on i -th output of circuit C . Actually, this function takes 1 value on Boolean vectors for that of which changing the value of variable v alters the value of i -th output of circuit C .

$$D_v f_i = f_i^{v=0} \oplus f_i^{v=1}. \quad (1)$$

Here $f_i^{v=0} = f_i(0, x_1, \dots, x_n)$, $f_i^{v=1} = f_i(1, x_1, \dots, x_n)$.

$$D_v f_i = f_i^{v=0} \overline{f_i^{v=1}} \vee \overline{f_i^{v=0}} f_i^{v=1}. \quad (2)$$

ROBDDs $R(f_i^{v=0})$, $R(f_i^{v=1})$, $\overline{R(f_i^{v=0})}$, $\overline{R(f_i^{v=1})}$ present functions $f_i^{v=0} = f_i(0, x_1, \dots, x_n)$, $f_i^{v=1} = f_i(1, x_1, \dots, x_n)$ and their inversions.

For getting observability function f^{obs} , for circuit C as a whole we use the following formula:

$$f^{obs} = \bigvee_{i=1}^{m_v} (D_v f_i) = \bigvee_{i=1}^{m_v} (f_i^{v=0} \overline{f_i^{v=1}} \vee \overline{f_i^{v=0}} f_i^{v=1}) \quad (3)$$

Here m_v is the number of outputs of circuit C connected with node v .

Note as $R(\phi)$ ROBDD representing the completely specified function ϕ that is implemented by sub-circuit C_v . Then on-set $M_1(f_v)$ and off-set $M_0(f_v)$ of incompletely specified Boolean function f_v corresponding to node v are represented by ROBDDs $R_1(v)$, $R_0(v)$, correspondingly:

$$R_1(v) = R(\phi)R^{obs} = R(\phi) \left[\bigvee_{i=1}^{m_v} (R(f_i^{v=0}) \overline{R(f_i^{v=1})}) \vee \overline{R(f_i^{v=0})} R(f_i^{v=1}) \right]$$

$$R_0(v) = \overline{R(\phi)R^{obs}} = \overline{R(\phi)} \left[\bigvee_{i=1}^{m_v} (R(f_i^{v=0}) \overline{R(f_i^{v=1})}) \vee \overline{R(f_i^{v=0})} R(f_i^{v=1}) \right]$$

ROBDDs $R_1(v)$, $R_0(v)$ are compact representations of incompletely specified function f_v .

When we have a set of fault nodes $V = \{v_1, \dots, v_q\}$, it is necessary to obtain incompletely specified Boolean function for each fault node v_i from V , representing it by the corresponding ROBDDs $R_1(v_i)$, $R_0(v_i)$.

IV. MAPPING AN INCOMPLETELY SPECIFIED BOOLEAN FUNCTION ON INTERNAL VARIABLES OF CIRCUIT C

We have got the incompletely specified Boolean function for node v depending on variables x_1, x_2, \dots, x_n . We suggest to map on-set and off-set of this function on internal variables u_1, \dots, u_m of circuit C . These internal variables at the same time are internal variables of sub-circuit C_v . A set of variables u_1, \dots, u_m may be chosen in the different way but sub-circuit with inputs u_1, \dots, u_m and output v together with sub-circuits with inputs x_1, x_2, \dots, x_n and outputs u_1, \dots, u_m comprise sub-circuit C_v .

Incompletely specified Boolean function for node v that depends on variables u_1, \dots, u_m is derived. Call this function as $f_v(u_1, \dots, u_m)$. It is implemented by fragment $C_v(u_1, \dots, u_m)$ of sub-circuit C_v with output node v and input nodes u_1, \dots, u_m . Note that the number m is not so much (about ten).

We suggest to apply the special ternary simulation [4] in order to find first two sets of ternary vectors on variables u_1, \dots, u_m that contain on-set and off-set of function $f_v(u_1, \dots, u_m)$.

Having got these vectors further we find on-set and off-set of incompletely specified function $f_v(u_1, \dots, u_m)$ using binary simulation on variables u_1, \dots, u_m and fragment $C_v(u_1, \dots, u_m)$.

A. Ternary simulation procedure

Execute ternary simulation for incompletely specified Boolean function f_v applying its representation by two ROBDDs $R_1(v)$, $R_0(v)$. For that a random approach is used [6].

Any product K originates the probability distribution of 1 value on input variables as follows. As we know, a product may be represented by the ternary vector (cube) with 1, 0, - (don't care) values of components. If variable x_i appears in product K without inversion (with inversion), the corresponding element in probability distribution takes 1(0) value. The same value takes this variable in ternary vector(cube). If variable x_i is absent in product K , the corresponding element in probability distribution takes $\frac{1}{2}$ value. In the ternary vector variable x_i takes value «-» (don't

care). For example, product $K = x_1 \overline{x_2} x_5$ is represented by ternary vector (cube) 10--1-. This product originates probability distribution $1 \ 0 \ \frac{1}{2} \ \frac{1}{2} \ 1 \ \frac{1}{2}$.

Let we have ROBDD R representing the Boolean function f and probability distribution $\rho(K)$ of 1 values of this function variables obtained from K . Note that K is generated by the certain path from the ROBDD R root till its 1 terminal node. It is possible to calculate probability $\rho(f)$ of 1 value of function f under probability distribution $\rho(K)$. For that in current internal node μ of ROBDD R we do the following.

Probability $\rho(\eta)$ of 1 value of Boolean function η , corresponding to ROBDD R internal node μ is calculated with using probabilities $p(\eta_{\mu}^{x_i=0})$, $p(\eta_{\mu}^{x_i=1})$ of 1 values of functions $\eta_{\mu}^{x_i=0}$ and $\eta_{\mu}^{x_i=1}$, corresponding to daughter nodes of node μ in the following way (node μ is marked by variable x_i): $p(\eta) = p(x_i)p(\eta_{\mu}^{x_i=1}) + p(\overline{x_i})p(\eta_{\mu}^{x_i=0})$.

Calculation of probability $\rho(f)$ is linear function of the number of internal nodes of ROBDD R .

If $\rho(f)$ is equal to 1, it means that product K is implicant of the function that is function f takes 1 value on all Boolean vectors turning product K into 1.

If $\rho(f)$ is equal to 0, it means that product K is implicant of inversion of this function that is function f takes 0 value on all Boolean vectors turning product K into 1.

Otherwise function f takes on Boolean vectors turning product K into 1 either 1 or 0 values.

If $\rho(f)$ is equal to 1 (0), function f takes 1 (0) value on ternary vector representing product K , otherwise function f takes don't care value noticed as «-» on this vector [4].

Let we have m ROBDDs $R(u_1), \dots, R(u_m)$ representing functions corresponding to sub-circuits with outputs u_1, \dots, u_m and inputs x_1, \dots, x_n and ternary vector α representing product K . It is necessary to map α into ternary vector β on variables u_1, \dots, u_m . For that we do the following.

Form probability distribution $\rho(\alpha)$ from α in above mentioned way.

Substitute probability distribution $\rho(\alpha)$ into ROBDDs $R(u_1), \dots, R(u_m)$ and get values of vector β components in above mentioned way.

Note that ternary vector β presents the values of functions implemented by the sub-circuits with outputs u_1, \dots, u_m and inputs x_1, \dots, x_n on ternary vector α .

B. Finding sets M_1^*, M_0^*

In order to map on-set $M_1(f_v)$ and off-set $M_0(f_v)$ of incompletely specified Boolean function f_v into on-set $M_1(f_v(u_1, \dots, u_m))$ and off-set $M_0(f_v(u_1, \dots, u_m))$ of incompletely specified Boolean function $f_v(u_1, \dots, u_m)$ we first derive M_1^*, M_0^* sets of ternary vectors (cubes) on variables u_1, \dots, u_m that contain Boolean vectors of sets $M_1(f_v(u_1, \dots, u_m)), M_0(f_v(u_1, \dots, u_m))$, correspondingly. We obtain M_1^*, M_0^* using ternary simulation described above.

We have ROBDD $R_1(v)$ representing $M_1(f_v)$. Each path from root of $R_1(v)$ till its 1 terminal node presents product K

(the ternary vector). For each ternary vector α derived from $R_1(v)$ we find ternary vector β . As a result we get set M_1^* . Set M_0^* is formed by the similar way from $R_0(v)$.

Obtain sets $M_1(f_v(u_1, \dots, u_m)), M_0(f_v(u_1, \dots, u_m))$ from sets M_1^*, M_0^* . Each ternary vector (cube) from M_1^* originates the corresponding set of Boolean vectors. The Boolean vectors from this set that turn sub-circuit C_v into 1 are included into $M_1(f_v(u_1, \dots, u_m))$. Similarly, the Boolean vector originated by a ternary vector (cube) from M_0^* is included into set $M_0(f_v(u_1, \dots, u_m))$ if this vector turns sub-circuit C_v into 0. We obtain sets $M_1(f_v(u_1, \dots, u_m)), M_0(f_v(u_1, \dots, u_m))$ using binary simulation of Boolean vectors originated by sets M_1^*, M_0^* .

If we have a set $V = \{v_1, \dots, v_q\}$ of fault nodes it is necessary to get sets $M_1(f_{v_j}(u_1, \dots, u_m))$, $M_0(f_{v_j}(u_1, \dots, u_m))$ for each node v_j from V . In this case there is a problem of finding the proper set u_1, \dots, u_m of internal variables for V but it is a subject of further study.

V. DERIVING MASKING CIRCUIT

When incompletely specified Boolean function is represented by two sets of Boolean vectors $M_1(f_v(u_1, \dots, u_m)), M_0(f_v(u_1, \dots, u_m))$ it is possible to apply ESPRESSO [7] system to get the corresponding Sum of Products (SoP) of completely specified Boolean function and then get patch circuit C_p using ABC system [8]. This patch circuit inputs are connected with internal nodes u_1, \dots, u_m of circuit C and its output is connected with internal nodes of circuit C that are fed by node v (Fig. 1).

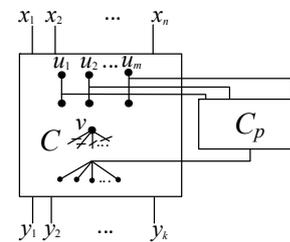


Fig. 1. Patching with using internal variables of circuit C

If patch circuit inputs are connected with circuit C inputs, we derive two sets of ternary vectors (cubes) from ROBDDs $R_1(v), R_0(v)$ to obtain SoP of completely specified Boolean function using ESPRESSO system and then we get patch sub-circuit using ABC system (Fig. 2).

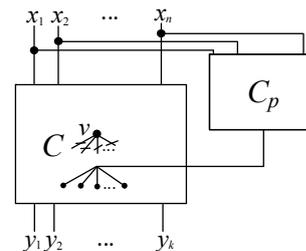


Fig. 2. Patching with using input variables of circuit C

If we have a set $V = \{v_1, \dots, v_q\}$ of fault nodes, we get patch circuit with q outputs. For that we obtain system of incompletely specified Boolean functions either on input variables of circuit C or internal variables of this circuit.

If we mask Trojan Circuits included into line, C_p output is connected with the only gate that is fed by this line.

VI. EXPERIMENTAL RESULTS

Experimental results are executed for circuits of system MCNC transformed into circuits consisting of gates with one

or two inputs by applying ABC system. Patch circuits are built for nodes with poor determined incompletely specified Boolean functions that is characterized by low values of observability [5]. For each circuit we choose several nodes. Experimental results are represented in Table 1. It is considered that only one node is fault. Overhead is ratio of the gate number of circuit C_p to gate number of circuit C (in percentage).

TABLE I. EXPERIMENTAL RESULTS FOR PATCH FUNCTIONS OF CIRCUIT C

Circuits C	The number of gates of circuit C	Chosen nodes	Observability of nodes	The number of gates of sub-circuit C_v	The number of C_p gates connected with C inputs	Overhead	The number of gates of sub-circuit $C_v(u_1, \dots, u_m)$	The number of C_p gates connected with C internal variables	Overhead
9symml	315	n124	0.126953	37	25	7,94%	10	11	3,49%
		n118	0.078125	32	21	6,67%	13	14	4,44%
		n147	0.0820313	33	20	6,35%	9	11	3,49%
alu4	1046	n620	0.0078125	301	10	0,96%	11	2	0,19%
		n618	0.0078125	300	9	0,86%	9	1	0,1%
		n731	0.0078125	285	20	1,91%	13	7	0,67%
C432	379	n115	0.0703947	133	52	13,72%	18	8	2,11%
		n177	0.101033	245	79	20,84%	15	16	4,22%
comp	158	n132	0.000244141	34	28	17,72%	11	4	1,06%
		n133	0.0020752	70	45	28,48%	13	4	1,06%
term1	420	n328	0.0621948	51	33	7,86%	8	15	3,57%
		n329	0.0625	61	33	7,86%	9	13	3,1%
		n250	0.0625	51	36	8,57%	9	9	2,14%

These experiments show that the suggested approach to masking faults (Trojan Circuits) may be effective for node v , if it is far from circuit C inputs. In this case we derive patch circuit that may be essentially simpler in comparison with duplication. We take in mind that circuit C_v ($C_v(u_1, \dots, u_m)$) with fault free gates may be applied as patch circuit but our aim to get simpler circuit if possible. Patch circuits connected with internal variables of circuit C provide, as a rule, several times less overhead compared with patch circuits connected with input variables of circuit C .

CONCLUSION

The new approach to patching is suggested. It is based on applying incompletely specified Boolean functions for fault nodes. The new algorithm of mapping incompletely specified Boolean functions depending on input variables of circuit C into its internal variables is developed. It is based on applying special ternary simulations on ROBDDs representing sub-circuits that outputs correspond to this internal variables but inputs are input variables of circuit C . A complexity of calculations connected with this simulation is linear function of the number of internal nodes of these ROBDDs. In Table 1 we show results of patching for one fault node. It is necessary to extend them for several fault nodes. When patch circuit inputs are connected with circuit inputs, we join incompletely specified Boolean functions of all fault nodes into system. Using incompletely specified Boolean functions on internal variables is more interesting problem in ECO practice. In this case it is necessary to solve the problem of choosing internal nodes that precede several fault nodes. We are going to solve this problem in future.

REFERENCES

- [1] A. Q. Dao, N.-Z. Lee, L.-C. Chen, M.P.-H. Lin, J.-H.R. Jiang, A. Mishchenko, and R. Brayton, "Efficient computation of ECO patch functions," in Proc. DAC, 2018.
- [2] A.-C. Cheng, H.-R. Jiang, and J.-Y. Jou, "Resource-aware functional ECO patch generation," in Proc. DATE, 2016.
- [3] S. Yamashita, H. Yoshida, and M. Fujita, "Increasing yield using partially-programmable circuits," in Workshop on Synthesis And System Integration of Mixed Information technologies (SASIMI), 2010, pp. 237–242.
- [4] H. Mangassarian, H. Yoshida, A. Veneris, S. Yamashita, and M. Fujita, "On error tolerance and Engineering Change with Partially Programmable Circuits," 2012, pp. 695–700.
- [5] Matrosova A., Ostanin S. Trojan Circuits Masking and Debugging of Combinational Circuits with LUT Insertion //2018 IEEE International Conference on Automation, Quality and Testing, Robotics. AQTR 2018 (THETA 21), 24-26 may 2018, Cluj-Napoca, Romania. [Cluj-Napoca], 2018. P. 462-467. 1 CD-R.
- [6] A. Matrosova, O. Goloubeva, S. Tsurikov, "On correction of the results of ternary simulation and preliminary estimation of the correction results", Proceedings of the 6-th Biennial Conference on Electronics and Microsystems Technology, Tallinn. 1998. P. 183-186
- [7] ESPRESSO: Logic Minimization Software (<http://ramos.elo.utfsm.cl/~lsb/elo211/aplicaciones/aplicaciones/espresso/ESPRESSO Logic Minimization Software.htm>).
- [8] ABC: A System for Sequential Synthesis and Verification (<https://people.eecs.berkeley.edu/~alanmi/abc/>).

Masking Robust Testable PDFs

Anzhela Matrosova
Institute of Applied Mathematics and
Computer Science
Tomsk State University
Tomsk, Russia
mau11@yandex.ru

Sergei Ostanin
Institute of Applied Mathematics and
Computer Science
Tomsk State University
Tomsk, Russia
sergeiostanin@yandex.ru

Semen Chernyshov
Institute of Applied Mathematics and
Computer Science
Tomsk State University
Tomsk, Russia
semen.cher@mail.ru

Abstract—High performance logical circuits are characterized by, first of all, their high operation speed determined by a clock frequency. The maximum possible clock frequency with proper functioning is one of the most important tasks in developing such circuits. Unfortunately, unpredictable delays may appear during functioning of high performance circuits. These delays decrease performance speed. They have to be detected. One of more effective model of delays is Path Delay Fault (PDF) model. There are two classes of PDFs: robust testable PDFs and non-robust testable PDFs. Detecting robust testable PDFs gives possibility to find fault paths precisely. This information may be applied by circuit developers to remove these delays in order to increase operation speed. As far as we know, they use physical methods for that. Here we for the first time suggest logical method of removing unpredictable delays in the frame of Engineering Change Ordering (ECO) technologies. Our approach is based on using compact representation of all test patterns v_2 (from test pairs (v_1, v_2)) detecting PDFs of the path considered. All test patterns are described as ROBDDs R_{rise} - for rising transitions of the path and R_{fall} - for falling transitions of the path. The ROBDDs are applied to design patch circuit. Experimental results showed that maximal lengths of paths in patch circuit are essentially shorter than maximal lengths of paths in the given circuit that are masked by the patch circuit. The more complicate the given circuit the smaller a portion of overhead in comparison with this circuit.

Keywords—combinational circuit, Reduced Ordered Binary Decision Diagram (ROBDD), Path Delay Fault (PDF), patch circuit.

I. INTRODUCTION

In high performance logical circuits may appear unpredictable delays that change correct functioning and decrease performance speed. One of more effective model of delays is path delay fault (PDF) model [1, 2]. To improve the reliability of such circuits is proposed to mask PDFs using patch circuits.

All PDFs are compounded from two classes: non-robust testable and robust testable faults [3]. If the test pair exists only when other circuit paths are fault free then a PDF is named non-robust testable. If there is a test pair so that the delay fault is detected on the path regardless of delay faults of other circuit paths then a PDF is named robust testable.

Note that if non-robust testable PDF is detected, we cannot exactly determine the path on that of which delay manifests itself but when detecting robust testable PDF we exactly know the fault path. In this paper we suggest masking delay faults of robust testable PDFs.

If (v_1, v_2) is a test pair for robust testable PDF then Boolean vector v_2 can be obtained as the test pattern for the literal constant fault of the Equivalent Normal Form (ENF) [4] extracted from a combinational circuit C (combinational

part C of a sequential circuit) [5]. Remind that path α of combinational circuit C is a sequence of logical elements in that of which the output of a previous element is the input of the subsequent element. One of inputs of the first element is one of inputs of circuit C and the output of the last element of the path is one of outputs of circuit C .

The literal 0 stuck-at fault corresponds to delay of rising transition of path α , the literal 1 stuck-at fault corresponds to delay of falling transition of path α . Here path α is in keeping with the literal. Note that the test pattern for stuck at 1 turns circuit C output corresponding to α into 0 and the test pattern for stuck-at 0 turns the same output into 1.

Boolean vector v_2 being the test pattern for 0(1) stuck-at fault of the ENF literal may belong to either a test pair for robust testable PDF of path α or non-robust testable PDF of this path. Some test patterns of the literal may originate test pairs only for non-robust testable PDF of path α . That is why when delay of path α is detected as robust testable we have to mask this delay on all test patterns for 0, 1 stuck-at faults of the corresponding literal. Only in that case we guarantee masking delay of path α on any test pair on that of which delay of path α may manifest itself both for rising and falling transitions. Masking is based on Engineering Change Ordering (ECO) technologies [6, 7]. In [8] all vectors v_2 for 0, 1 stuck-at faults of the literal corresponding to path α are derived. They are compactly represented by the Boolean difference D_{path} .

In section II all test patterns v_2 for rising and falling transitions are obtained and compactly represented by the corresponding two ROBDDs. In section III the method of deriving patch circuit that masks robust testable PDFs is suggested. Experimental results are given in section IV.

II. DERIVING ALL TEST PATTERNS FOR RISING AND FALLING TRANSITIONS

Let α be a path of combinational circuit C (it may be a combinational part of a sequential circuit.) with inputs x_1, \dots, x_n that is represented by sequence of symbols: $x_i, u_1, u_2, \dots, u_{(r-1)}, u_r$. Here x_i is the beginning of the path α (circuit C input), $u_1, u_2, \dots, u_{(r-1)}, u_r$ are outputs of the path gates, r is the length of path α . Output u_r is one of the circuit C output.

Let $u_i, u_{(i-1)}$ be outputs of neighbor gates of path α . Consider sub-circuit C_{u_i} of circuit C . The sub-circuit inputs are x_1, \dots, x_n and $u_{(i-1)}$ where $u_{(i-1)}$ is also input variable of gate with output u_i .

Boolean difference for the function realized by sub-circuit C_{u_i} with respect to variable $u_{(i-1)}$ is called $D_{u_i} / D_{u_{(i-1)}}$

To find $D_{u_i} / D_{u_{(i-1)}}$ we have to execute the following steps.

1. Obtain $R(f_{u_i})$ presenting the function of sub-circuit C_{u_i} . The ROBDD depends on variables $x_1, \dots, x_n, u_{(i-1)}$. Variable $u_{(i-1)}$ is the first variable when applying the Shannon decomposition when ROBDD $R(f_{u_i})$ is derived.

2. Two ROBDDs $R(f_{u_i}^{u_{(i-1)}=1})$ and $R(f_{u_i}^{u_{(i-1)}=0})$ from $R(f_{u_i})$ are obtained. That roots are children nodes of ROBDD $R(f_{u_i})$ root. These ROBDDs realize functions that are derived from function f_{u_i} by changing variable $u_{(i-1)}$ for constant 1, 0, correspondingly.

3. Conjunctions $\bar{f}_{u_i}^{u_{(i-1)}=1}$, $f_{u_i}^{u_{(i-1)}=0}$, and $\bar{f}_{u_i}^{u_{(i-1)}=0}$, $f_{u_i}^{u_{(i-1)}=1}$ are implemented and results are merged being presented by ROBDD $R(D_{u_i} / D_{u_{(i-1)}})$:

$$R(D_{u_i} / D_{u_{(i-1)}}) = R(f_{u_i}^{u_{(i-1)}=0}) \& R(\bar{f}_{u_i}^{u_{(i-1)}=1}) \vee R(f_{u_i}^{u_{(i-1)}=1}) \& R(\bar{f}_{u_i}^{u_{(i-1)}=0}).$$

Note that ROBDD operations are characterized by a polynomial complexity.

4. Implement conjunctions:

$$R(D_{path}) = R(D_{u_r} / D_{u_{(r-1)}}) \& R(D_{u_{(r-1)}} / D_{u_{(r-2)}}) \& \dots \& R(D_{u_2} / D_{u_1}) \& R(D_{u_1} / D_{x_i}).$$

As a result we get ROBDD $R(D_{path})$ representing Boolean difference D_{path} for path α .

Boolean vector μ turns $R(D_{path})$ into 1 and turns circuit C into 0, then it is vector v_2 for falling transition. If vector μ turning $R(D_{path})$ into 1 and circuit C into 1 is vector v_2 for rising transition [5].

Thus, ROBDD $R(D_{path})$ represents test patterns both for falling and rising transitions. It is necessary to divide them.

ROBDD for falling transition call as R_{fall} :

$$R_{fall} = R(D_{path}) \& \bar{x}_i (R(D_{path}) \& x_i).$$

ROBDD for rising transition call as R_{rise} :

$$R_{rise} = R(D_{path}) \& x_i (R(D_{path}) \& \bar{x}_i).$$

Illustrate this procedure by example.

Note that forming R_{rise} we multiply $R(D_{path})$ and input variable x_i that inversion sign coincides with the inversion sign of the literal, corresponding to path α in ENF. Forming R_{fall} we apply variable x_i with opposite inversion sign. Remind that the inversion sign of the literal corresponding to path α is determined by the number of inversion gates along path α .

We have a combinational circuit (Fig. 1).

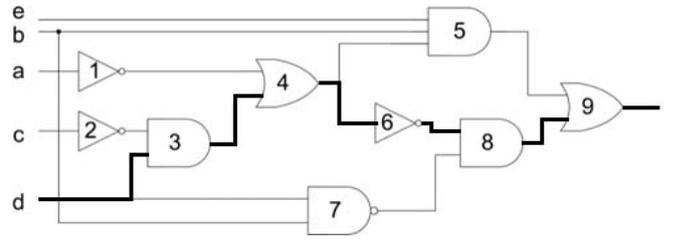


Fig. 1. Combinational circuit C

For example, path α includes gates with numbers 3, 4, 6, 8, 9. The beginning of the path is marked by variable d . Path α is presented by the sequence: $d, u_3, u_4, u_6, u_8, u_9$.

Derive D_{path} for α representing results by sum of products (SoP) for simplicity.

$$D_{u_9} / D_{u_8} = (u_5 \vee (u_8 = 1)) \oplus (u_5 \vee (u_8 = 0)) = \bar{u}_5 = \bar{b} \vee \bar{e} \vee ac \vee ad.$$

$$D_{u_8} / D_{u_6} = (u_6 = 1) \& u_7 \oplus (u_6 = 0) \& u_7 = u_7 = \bar{b} \vee \bar{d}.$$

$$D_{u_4} / D_{u_3} = (\bar{a} \vee 1) \oplus (\bar{a} \vee 0) = a.$$

$$D_{u_3} / D_d = (\bar{c} \& 1) \oplus (\bar{c} \& 0) = \bar{c}.$$

$$D_{path} = (\bar{b} \vee \bar{e} \vee ac \vee ad) \& (\bar{b} \vee \bar{d}) \& a \& \bar{c} = \bar{a}\bar{c}\bar{b} \vee \bar{a}\bar{c}\bar{d}.$$

Obtain SoP_{rise} , presenting all vectors v_2 for rising transition and SoP_{fall} presenting all vectors v_2 for falling transition.

$$SoP_{rise} = (\bar{a}\bar{c}\bar{b} \vee \bar{a}\bar{c}\bar{d}) \& \bar{d} = \bar{a}\bar{c}\bar{d}.$$

$$SoP_{fall} = (\bar{a}\bar{c}\bar{b} \vee \bar{a}\bar{c}\bar{d}) \& d = \bar{a}\bar{c}\bar{b}d.$$

Note that we have to mask rising and falling transitions of path α using test patterns represented by ROOBD R_{rise} and ROOBD R_{fall} . Test patterns from R_{rise} detect 0 stuck-at fault of ENF literal corresponding to path α and when delay on this path takes place we observe on the corresponding output of circuit C 0 value instead of 1 value. Test patterns from R_{fall} detect 1 stuck-at fault of ENF literal corresponding to path α and when delay on this path takes place we observe on the corresponding output of circuit C 1 value instead of 0 value.

Denote the sub-circuit containing path α as C_s (sub-circuit of specification) if path α is fault free. If we detect PDFs for rising and falling transitions on path α , this sub-circuit is denoted as C_i (implementing sub-circuit).

Take into account that we consider the longest paths of circuit C as delay faults, as a rule, appear on such paths. We suppose that ROBDDs R_{rise} , R_{fall} for long paths are rather simple and originate not so much cubes [9]. Remind that any cube is generated by the ROBDD path that runs from the ROBDD root till its 1 terminal node. It means that we may use the approach suggested by us in [10] that is oriented to slight difference between sub-circuit of specification C_s and implementing sub-circuit C_i .

It is suggested that we have spare cells for masking circuit. In this paper the masking circuit (patch circuit) is connected with circuit C_i as follows. Inputs of the patch

circuit are connected with inputs of circuit C_i and the output of the patch circuit is connected with the output of C_i .

III. MASKING DELAYS OF ROBUST TESTABLE PDFS

Thus we have sub-circuit C_i . It implements function f_i . In this sub-circuit we detected delay of path α for rising and falling transitions. Execute the following steps to mask this delay.

1. Form List L of cubes (ternary vectors) with their 0(1) characteristics applying ROBDDs R_{rise} , R_{fall} . If the cube characteristic is 0(1), it means that on any Boolean vector of this cube output of C_i is equal to 0(1) but output of C_s takes opposite 1(0) value.

2. Divide list L into L_0 and L_1 .

Include into L_0 cubes with 0 characteristic originated by R_{rise} and into L_1 - cubes with 1 characteristic originated by R_{fall} .

3. List L_0 represents on-set $M_{C_0}^1$ of the function realized by circuit C_0 that is a sub-circuit of the patch circuit and L_1 represents on-set $M_{C_1}^1$ of the function realized by circuit C_1 that is another sub-circuit of the patch circuit. It means that $M_{C_0}^1$ consists of cubes extracted from R_{rise} and $M_{C_1}^1$ consists of cubes extracted from R_{fall} .

Note the following: to get circuit C_s we need to extend on-set of C_i using $M_{C_0}^1$, and exclude on-set $M_{C_1}^1$. It means that $M_{C_s}^1 = M_{C_0}^1 \cup (M_{C_i}^1 / M_{C_1}^1)$. In other words, $M_{C_s}^1 = M_{C_0}^1 \cup (M_{C_i}^1 \cap \bar{M}_{C_1}^1)$.

This situation is illustrated by Venn diagram (Fig. 2).

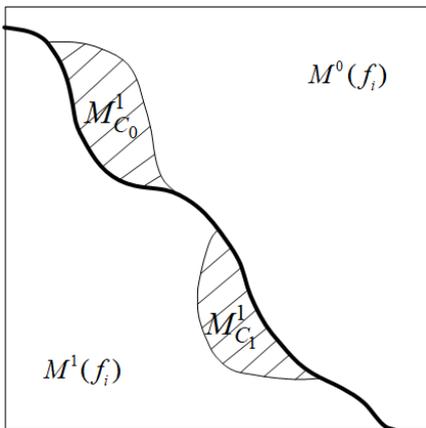


Fig. 2. Correction areas of function f_i

On this figure function f_i is represented by its on-set $M^1(f_i)$ and off-set $M^0(f_i)$. These sets are parted by thick line. Dashed areas (correction areas) represent $M_{C_1}^1$ among $M^1(f_i)$ and $M_{C_0}^1$ among $M^0(f_i)$.

Circuit C_s is superposition of circuits C_0 , C_1 and C_i . It is represented by Fig. 3.

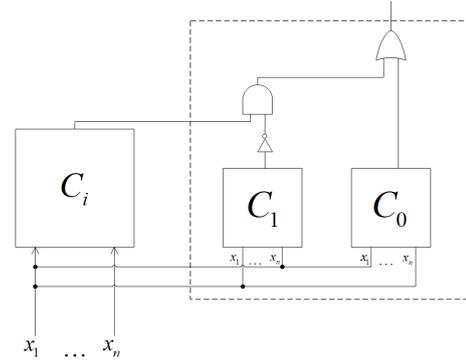


Fig. 3. Circuit C_s

In Fig. 3 the patching sub-circuit is outlined by dashed line.

Note that path α may be fault free either for rising or falling transition. In this case, we have smaller patch circuit originated by one of ROBDDs R_{rise} , R_{fall} .

If we detect robust testable PDFs for t paths, we have to derive sets L_0 , L_1 for each fault path and design circuits C_0 , C_1 with r outputs. Here r is not more t as different fault paths may be ended on the same output of circuit C_i .

IV. EXPERIMENTAL RESULTS

Circuits from ISCAS'89 are considered. First only delays of one path for rising and falling transitions are masked with the patch circuit. For each circuit from ISCAS'89 several the longest paths are examined (Table I). Patch circuits are designed with using ABC system [11]. Two level synthesis method is implied. In Table II three the longest paths of the circuit are masked by one patch circuit.

In the second column of Tables I, II the names of circuits are given, in the third and the fourth columns - the numbers of circuit elements and the lengths of paths of circuits are shown. In the fifth and sixth columns of these tables the numbers of elements and the maximal lengths of paths of patch circuits are given. In the seventh column the overhead (in percentage) is represented.

We see that the more complicate circuit C_i the less overhead in percentages for masking delays. It is important that the maximal path lengths in patch circuits are essentially shorter the maximal path lengths in the given combinational circuit (circuit C_i).

CONCLUSION

As far as we know the approach to masking robust testable PDFs is suggested for the first time. It is based on using compact representation of test patterns v_2 of test pairs (v_1, v_2) for rising and falling transitions of PDFs of the path considered by two ROBDDs. These test patterns are applied for design of the patch circuit. Experimental results show that maximal lengths of paths in patch circuit are essentially shorter than maximal lengths of the paths that delays are masked by the patch circuit. Taking into consideration the overhead a circuit developer may choose either decreasing speed functioning or increasing overhead.

TABLE I. MASKING DELAYS OF ONE PATH

№	Circuit name	The number of circuit elements	Maximal length of the path of the given circuit	The number of elements of patch circuit	Maximal paths length of patch circuit	Overhead in percentages
1	s208	96	14	15	4	15.6
2	s208	96	12	13	3	13.5
3	s298	119	8	9	3	7.6
4	s298	119	8	12	3	10
5	s298	119	8	12	3	10
6	s298	119	8	11	3	9.2
7	s382	158	9	21	4	13.8
8	s382	158	9	21	4	13.3
9	s382	158	9	23	4	14.6
10	s386	159	11	18	4	11.4
11	s386	159	11	16	4	10
12	s386	159	11	16	4	10
13	s444	181	11	23	4	12.7
14	s444	181	11	21	4	11.6
15	s953	395	16	30	4	7.6
16	s953	395	16	30	4	7.6
17	s953	395	15	20	4	5
18	s953	395	15	25	4	6.3

TABLE II. MASKING DELAYS OF THREE CIRCUIT PATHS

№	Circuit names	The number of circuit elements	Maximal length of the paths of the given circuit	The number of gates of patch circuit	Maximal paths length of patch circuit	Overhead in percentages
1	s208	96	14	19	4	19.8
2	s298	119	8	18	4	15.1
3	s344	160	20	30	6	18.7
4	s382	158	9	18	6	11.4
5	s386	159	11	24	5	15.1
6	s444	181	11	18	5	9.9
7	s953	395	16	37	5	9.4

REFERENCES

- [1] G.L. Smith, "Model for delay faults based upon paths," in Proc. of the International Test Conference, 1985, pp. 342–349.
- [2] C.J. Lin, S.M. Reddy, "On delay fault testing in logic circuits," in IEEE Trans. Comput-Aided Design Integr. Circuits Syst., 6(5), 1987, pp. 694–703.
- [3] A. Krstic, K.T. Cheng, "Delay fault testing for VLSI circuits," Kluwer, Boston, 1998.
- [4] D.B. Armstrong, "On Finding a Nearly Minimal Set of Fault Detection Tests for Combinational Logic Nets," IEEE Transactions on Electronic Computers EC-15, no. 1 (February 1966), pp. 66–73.
- [5] A. Matrosova, V. Lipsky, A.Melnikov, V. Singh, "Path delay faults and ENF," IEEE East-West Design&Test Symposium. St. Petersburg: IEEE, 2010, pp. 164-167.
- [6] A.Q. Dao, N.-Z. Lee, L.-C. Chen, M.P.-H. Lin, J.-H.R. Jiang, A. Mishchenko, and R. Brayton, "Efficient computation of ECO patch functions," in Proc. DAC, 2018.
- [7] A.-C. Cheng, H.-R. Jiang, and J.-Y. Jou, "Resource-aware functional ECO patch generation," in Proc. DATE, 2016.
- [8] A.Yu. Matrosova, V.V. Andreeva, E.A. Nikolaeva, "Finding Test Pairs for PDFs in Logic Circuits Based on Using Operations on ROBDDs," Russian Physics Journal. 2018. Vol. 61, № 5. pp. 994-999.
- [9] A.Yu. Matrosova, V.V. Andreeva, Tychinskiy V.Z., Goshin G.G. "Applying ROBDDs for delay testing of logical circuits," Izvestia vyzov. Physics. 2019.v. 62, № 5. pp. 86-94.
- [10] A. Matrosova, S. Chernyshov, G. Goshin, D. Kudin, "Forming Patch Functions and Combinational Circuit Rectification," Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2018), Kazan, 14-17 september 2018. [S. 1.], 2018. pp. 726-730.
- [11] ABC: A System for Sequential Synthesis and Verification — URL: <https://people.eecs.berkeley.edu/~alanmi/abc/>.

Associative Processors: Application, Operation, Implementation Problems

Egor Kuzmin

*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)
Moscow, Russia
kuzminen@yandex.ru*

Nikolay Levchenko

*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)
Moscow, Russia
nick@ippm.ru*

Anatoly Okunev

*Department of high-performance
microelectronic computing systems
Federal State-Funded Institution of
Science Institute for Design Problems
in Microelectronics of Russian
Academy of Sciences (IPPM RAS)
Moscow, Russia
oku@ippm.ru*

Abstract—In this article, we looked at many different architectures of associative computing systems and tried to find common features and differences among them. The various fields of application of associative processors, the problems associated with their use, as well as the specifics of their functioning and the relevance of their application were discussed. Methods for improving the energy efficiency of the computing system, which are used in the development of modern associative processors were analyzed and studied. These methods were compared with those used in the development of the PDCS “Buran” for solving the problem of improving the system’s energy efficiency.

Keywords—*associative processors, methods of power optimization, parallel dataflow computing system*

I. INTRODUCTION

With an increase in the volume of processed data, higher demands are placed on computing systems (CS) related to the possibility of parallelizing calculations between execution units of the CS.

The traditional von Neumann architecture of the CS was designed as architecture with sequential execution of commands. Von Neumann architecture is characterized by two principles: the principle of programmed management of program execution and the principle of the program storage in the memory. This architecture uses a command counter, which just limits the flow of commands that come to be executed, replacing them with a sequential analysis of incoming commands. The limited parallelism in exchanges between the cores and the memory does not allow increasing the productivity in proportion to the number of cores, productivity growth stops for many tasks already on several dozen cores. [1]

In modern CS, multiprocessor (multi-core) hybrid architectures of the MIMD / SIMD class are used, which allow speeding up the task performance due to its splitting into sequential and parallel sections.

Moreover, parallelism in modern general-purpose (GPP) and special-purpose (ASP) processors can be maintained at all levels (micro-level parallelism, command level parallelism, thread level parallelism, task level parallelism). At the same time, in addition to traditional von Neumann cores, modern processors also use various kinds of accelerators. [2, 3]

One of the computing devices with an architecture that is different from the sequential von-Neumann architecture, and

which allows to increase the level of parallelism of the entire system, is an associative processor.

Systems using an associative processor in their composition will allow efficient use of mass fine-grained data parallelism. For many actual tasks, using an associative processor will give advantages over solutions obtained on other computers using GPP or ASP.

In recent years, when developing architectures of high-performance computing systems, there has been a tendency that when choosing the architecture for the cores that make up the system, preference is increasingly given to simpler cores. The choice of simple cores is not accidental, since it can allow achieving a higher index of density on the chip, as well as increasing the level of energy efficiency of the entire system. In this regard, the use of an associative processor is the next step after the appearance of a GPU calculator in the direction of increasing the number of computing cores in the system and simplifying the structure of an individual core, since in an associative processor each row of an associative drive array can be the simplest computing element.

Systems using an associative processor in their composition will allow efficient use of mass fine-grained data parallelism. For many actual tasks, using an associative processor will give advantages over solutions obtained on other computers using GPP or ASP.

Associative processors (used for high-performance computing) were used, for example, in the Maxeler system and in the parallel dataflow computing system (PDCS) “Buran” [4] developed in the IPPM RAS. Both of these architectures implement a computational model with dataflow control (dataflow), the difference is that Maxeler uses the principles of static dataflow, and the PDCS “Buran” supports a streaming computation model with a dynamically generated context [5].

II. FIELDS OF APPLICATION OF ASSOCIATIVE PROCESSORS AND THE SPECIFICS OF THEIR FUNCTIONING

The associative computer system used to look like a simple array of associative memory, implemented as a hash table (LUT-table). Such associative memory has found its application in TLB-buffers of processors, in routers, as well as in computing modules of high-performance computing systems to reduce redundant computing [6, 7, 8]. An associative memory can be either fully associative (when the comparison logic is present in each memory cell and the search is performed simultaneously on all words (rows of memory array) and data bits (columns of memory array) and

partially associative (the comparison logic is at the output of the memory, and not in each word; therefore, only one bit-slice bit column is loaded into it per clock cycle). A system consisting of an associative memory array supports only read and search operations on an array with previously stored information (for example, in routers).

Subsequently, with the development of the promising direction of “computation in memory” (PiM) [9, 10] and the emergence of all the new requirements necessary for solving applications, an apparatus was added to the array of associative memory, which allows for the implementation of basic arithmetic and logical operations, as well as complex associative operations. Search, which allowed the use of such associative systems as SIMD accelerators [4, 11] and led to the emergence of various full-fledged associative computing systems (STARAN, MACS, FPS “Buran”, Maxeler, etc.). For example, to bypass the disadvantage associated with the need to write words to memory sequentially (which reduces read / write time from / to memory), in [11] it was proposed to use part of the associative memory array as an input / output buffer. Also, for solving some applications, local addition is required within the group of CAM strings (for example, for the FIR filter) or vector calculations (which contain addition, multiplication, subtraction) simultaneously on two dimensions of the PA array, use a two-dimensional associative processor.

The associative processor supports three types of operations: compare-read, search (and support for complex multiple search is possible), compare-write.

The computation process in an associative processor can be represented as the assignment of a function to work with an array of stored data. Associated with this is the multifunctionality of the associative processor. (The computation process in associative processor can be represented as the assignment of a function to work with an array of stored data. That’s mean that the operations on the APs are performed by applying the truth table of the function in an ordered sequence to the CAM. The multifunctionality of the associative processor is associated with this feature.) The specifics of the operation and the algorithm of the associative processor (when performing matrix multiplication) were described in more detail in [12, 13, 14].

Such versatility and the ability to select data dimensions for each computational element by assigning a function to a part of an arbitrary data array, leads to the fact that a large number of different types of tasks (for example, associated with the implementation of machine learning algorithms and complex search algorithms) are much better scaled. It also makes the AP suitable for a large number of applications containing a huge amount of data.

Despite the fact that the principles of operation of associative memory were developed quite a long time ago and were often used in various nodes or blocks of computing systems, during this time a generalized standard of the architecture of an associative processor was not developed. The absence of a standard encourages developers to either release associative computing systems designed to solve highly specialized tasks, or try to optimize existing microarchitecture of associative processors for these tasks.

For example, for neuromorphic processors [15], the organization of memory in the form of parallel columns is typical, that is, data words are stored in the memory array as

an array column, the same for DIMA (deep in memory architecture) architecture [16].

However, most architectures of associative processors have much in common in their structures [17-22].

The basic blocks included in the structure of an associative processor are as follows:

- array of associative memory, each line of which is connected to a sensing amplifier;
- tag register or array of tags for multiple response [21, 22];
- key registers for writing the key value (which will then be searched) and a mask that ensures a bit match for any value;
- command memory and a controller that generates the required key and mask values for the corresponding command;
- interconnect matrix or electronic switch matrix that allows associative processor strings to interact in parallel with other associative processors. The interaction can be carried out both at the bit level and at the word level;
- registers storing values of current and previous matches.

The use of associative processors allows the following:

- Hardware implementation of associative systems. This includes the implementation of pulsed, convolutional, and other neural networks [15, 16], as well as the hardware implementation of deep machine learning algorithms and systems designed to solve specific problems.
- Achieve the advantages (performance, energy efficiency, etc.) when solving problems compared to computers with von Neumann architecture. This is due to the better scalability of the problems solved, supported by the system with an associative computational model.

Unconventional computing models (for example, dataflow computing model) are increasingly having an impact on the choice of an associative processor as an alternative to CPU and GPU calculators. This approach allows you to effectively use the system with associative processors on the set of tasks of different classes, namely:

- high-performance parallel computing;
- tasks with a matrix data structure (recognition and analysis of images, scenes and situations, image processing and radar information, speech recognition and synthesis);
- parallel processing of fuzzy information;
- hardware implementation of various functions of operating systems (cache-memory, search, sorting, packaging, data protection);
- tasks of machine translation;
- management tasks and quick search in databases.

The support of the content addressable memory in associative computing systems makes its usage the most attractive for tasks requiring fast (multiple) search over the entire data array, and the algorithm implies frequent memory access.

The following are the main areas of application of associative processors (AP):

- use of associative processors as a SIMD accelerator in hybrid computing systems;
- implementation on the associative processors of deep machine learning algorithms (deep ML) and the use of AP in decision support systems;
- use of associative processors in the construction of associative computer systems for solving highly specialized tasks;
- use of associative processors as a search module for working with unordered databases.

III. CHALLENGES FACING THE DEVELOPERS OF ASSOCIATIVE PROCESSORS

When designing architecture of any computing system, the developer always faces the search for a solution to the following fundamental problems:

- reduction of space occupied by the computing system;
- improving the energy efficiency of the entire system;
- increase system performance.

The solution to each of these problems is related to others, therefore, improving the performance of the system as a whole will be to find the optimal balance in solving these three tasks, while finding balance, it is necessary to take into account that the system characteristics (area, performance, energy efficiency) improved in within the framework of solving each problem, have a strong influence on each other. The search for a solution can occur at various hierarchical levels of the CS, ranging from the software and operating system used by the CS to the physical level, i.e. the material from which the components of the CS are made.

One of the determining factors that is taken into account when searching for the optimal solution of the three listed problems is the scope of the developed associative processor, i.e. what types of tasks will the workload of the CS consist of.

For example, for tasks related to the processing and analysis of non-numeric data of large volume, the use of AP assumes taking advantages over the GPU (and CPU). For this reason, as well as due to a higher degree of scalability of the AP compared to the GPU [23] (Fig. 1). The most obvious solution for the developer of the AP microarchitecture is the choice in favor of the maximum size of the associative array.

Since there is a direct relationship between the size of the associative memory array and the energy consumed by the associative processor, associative processor developers have directed efforts to create the most energy-efficient associative memory and improve search algorithms. The next chapter will present some of the techniques for reducing energy consumption used by associative processor developers.



Fig. 1. Comparison of the coverage of various tasks of various kinds of calculators (presentation of the company GSI Technologies [23])

IV. METHODS OF REDUCING POWER CONSUMPTION, USED BY DEVELOPERS OF ASSOCIATIVE PROCESSORS

Another surge of interest in the development of new architectures of associative processors occurred due to the growing needs in the new paradigm of “computations inside the memory” [10, 16], which is increasingly used as a solution that eliminates constraints peculiar to the background-Neumann architecture. As an additional stimulating factor for research, it is necessary to note the search for solutions (when developing CS architecture) among different parallel SIMD CS with “PiM”, which would optimally satisfy various system parameters (for example, the possibility of system scalability, structure and dimensionality of data used in operations, support for the level of fine-grained data parallelism, etc.) necessary to be able to solve the problem.

It is obvious that the emergence of new technologies of non-volatile memory served as a trigger to the explosive growth of the emergence of new APR-s architectures. The properties of various types of non-volatile memory (STT-RAM, MRAM, ReRAM), such as: memory cell size, density on a chip, compatibility with SoC, power consumption required for read and write operations, and the absence of leakage currents) allowed it to take DRAM and Flash memory niche and get solutions to some of the challenges faced by the SRAM-based APR developers from the first wave of research. CMOS-based SRAM memory in comparison still has the highest performance, has high energy efficiency when performing operations, but nevertheless it has significant drawbacks: the enormous size of the PL and the high rate of energy loss in the static mode of operation (due to charge leakage currents in AP array). These shortcomings were the reason for shifting the problems of developing associative aircraft to the search for the most energy efficient solution. A comparison of different memory technologies is presented in Fig. 2.

The emergence of non-volatile memory, allowed a new look at the solution of the challenges facing the first developers of associative computing systems, - increasing energy and area-efficiency. It is obvious that many of the considered methods of increasing the energy efficiency of the APR will be focused on taking advantage of the sharing of various memory modules: SRAM memory of a small amount implemented using CMOS technology (or STT-RAM), and nonvolatile memory of huge volume, used as the main data memory. And for sharing, several solutions are proposed at once:

	SRAM	DRAM	NOR	NAND	MRAM	STT-RAM	R-RAM
Data Retention	N	N	Y	Y	Y	Y	Y
Memory Cell Factor (F ²)	50-120	6-10	10	2-5	16-40	4-20	<4
Read Time (ns)	1	30	10	50	3-20	2-20	<50
Write /Erase Time (ns)	1	50	105-10 ⁷	10 ⁶ -10 ⁹	3-20	2-20	<100
Endurance	10 ¹⁶	10 ¹⁶	10 ⁷	10 ⁵	10 ¹⁵	10 ¹⁵	10 ¹⁵
Power Consumption – Read/Write	Low	Low	High	High	Med/High	Low	Low
Power Consumption – Other than R/W	Leakage Current	Refresh Power	None	None	None	None	None
Embedded/SoC Friendly	Y	N (Thermal)	N (Thermal)	N (Thermal)	Y	Y	Y

Fig. 2. Comparison of different types of memory

- SRAM (CMOS-based or STT-RAM based) memory stores data of a small amount, which is an address pointer to the main bulk data store. Such a solution was implemented in the development of one of the modules of the SATS “Buran” - an associative memory module (MAP) [24]. Also, this solution is often used to reduce redundant computations, when implementing associative memory within the architecture of a CPU or GPU pipeline, if a pipeline coincides, the conveyor is interrupted, and the output comes from the corresponding memory that stores ready-made answers; associative memory [6,7,8].
- In [25, 26], nonvolatile memory is proposed to be used in the Analog Matching-Cell module, to find similarity, by analogy, between the input vector and the template vectors loaded into the module from the SRAM memory via the DAC. In this case, the contour that calculates the address of the most matched pattern vector is implemented using CMOS technology. Further development of these works led to the emergence of several DIMA architectures [16].

All energy efficiency enhancement technicians of AP are usually:

- introducing a hierarchy into the data memory structure;
- segmentation of the PA array;
- the use of approximate calculations.

Introducing a hierarchy into the data memory structure, by dividing the entire system memory into several levels and selecting the best combinations of memory modules implemented using different technologies (SRAM STT-RAM MRAM etc.), based on their physical characteristics and advantages (storage density / cost search / write operations, etc.) that can be obtained from their use.

Segmentation of the PA array in order to allow parallel execution of various operations and to reduce the power of unused parts of the memory (Low activity) and to reduce the number of redundant calculations. Achieved by selective comparison method (selective compare = selective pre-charge + selective evaluate) [8, 27, 28] by splitting the search operation into several components - consecutive [8, 28] or parallel [27]. The articles [8, 28] presented a phased search in which a search word of a large length of N-bits is divided into several (m) parts, after which a sequential search is performed on each of them. In case of a mismatch, at some stage of the search, the further search does not continue,

which leads to energy savings and the ability to use long search words. When the search operation is divided into several successive stages, the delay associated with the search increases and the average processor idle time decreases. On the other hand, a parallel search, for which the search operation consists of two stages, will require the presence of a logical “OR” contour to find the string corresponding to all intermediate matches, which will lead to additional energy and space costs.

The downside of the benefits derived from using the selective comparison technique is the additional hardware costs in the form of two two-input logic “I” elements and one SRAM cell [29]. Another technique considered in the same paper [29] is the modified truth table method. For some operations (multiplication and absolute value) supported by the associative processor, the additional SRAM cell used in the implementation of the selective comparison method may keep the value longer than a single pass through the truth table.

The selective comparison method is an optimization at the architectural level, and the modified table method is an optimization at the instruction level. You can use both of the methods discussed separately or together.

The use of approximate calculations (Approximate Computing). This solution is suitable only for tasks that are tolerant of errors and allow for some level of inaccuracy of output results (problems of DSP, pattern recognition, text, etc.). Approximate calculations can be presented both at the logic circuit level and at the algorithmic-architectural levels:

- At the level of the electrical circuit, the most common method is to develop a functionally identical, but at the same time approximate electrical circuit for arithmetic functions.
- The approximation used at the algorithmic-architectural level is aimed at supporting the most significant components of the system to the detriment of less significant ones. For example, a computer program can ignore some iterations in the least significant cycles or skip a few memory accesses in the least important part of the calculations.

In an associative processor, approximations can also be represented at various levels:

- at the level of the logical contour - approximation is achieved by scaling the parameters of memory cells, in which, either the least significant columns of the word are supplied with lower supply voltages (for an associative memory array implemented using CMOS technology) [16, 25, 26], or the lower and upper resistances of the memristor are set (for an array of associative memory implemented on memristors) [27]. This method is called the voltage scaling method (VOS Voltage Overscaling). In an associative memory based on nonvolatile memory devices, the voltage scaling method can control the search energy.
- at the algorithmic-architectural level - approximation is achieved by “trimming” bits (bit trimming), which does not use columns corresponding to the least significant bits. In our sun, this can be represented by a masking operation.

When searching for solutions to improve the energy efficiency of the associative memory and the associative processor, a lot of research was conducted [6, 7, 8, 13, 27, 29, 30, 31], which proposed various solutions. Many of these solutions have already been implemented during the development of the PDCS Buran system [24, 28].

In modern studies related to the search for solutions to improve energy efficiency, isolated cases are considered, and in the system developed by us PPSU (at the same time) uses the combination of the majority of the considered solutions. Developers expect to benefit from the synergy of various solutions that should be expressed in increasing the level of concurrency of the CS (level of parallelization of the computational load) and in increasing the level of energy efficiency of the entire CS.

V. CONCLUSION

Since the possibilities of scaling CMOS technologies will be exhausted in the near future, developers of computing systems will need to concentrate on increasing the efficiency of calculations, not by improving the characteristics of transistors, but by increasing the level of parallelism of the entire system. An associative processor, being a SIMD calculator, makes it possible to achieve a high level of data processing parallelism due to its matrix structure (grid of PEs included in the AP) and a completely different (different from the address method) associative data processing method, in which the data is sampled by content (based on a match with the word search). In some APs, both data parallelism and command parallelism can be supported simultaneously.

The associative processor is the next step in development from CPU to GPU. The associative processor has a PE's (Processing Elements) with a simpler micro-architecture, compared with the computational elements that make up the modern GPGPU. Therefore, a higher level of parallelism is available to the associative processor, since a larger number of computational elements can be located on a single chip.

When comparing with other massively parallel CS, the associative processor wins (has a performance advantage) when solving problems containing many operations with unordered data and memory access operations, since the search operation for an n-bit word will take 1 tact (for a fully parallel AP) and n-cycles (for a bitwise-serial AP), which is significantly less than the results of a CS that support address data processing. Moreover, the associative processor, being an implementation of the concept of computations inside memory, implying the presence of elementary logic in the memory array for data processing through simple arithmetic and logical operations, allows one to get rid of the complex hierarchy of memory characteristic of for PDCS, and also from need for data transmission between EU (execution Unit) and memory.

The following areas are considered as directions for future research in the project to create the PDCS "Buran": building a specialized calculator based on an associative processor that supports the principles of a streaming model of calculations with a dynamically generated context; research and implementation of additional benefits derived from the use of an associative processor as a matching processor in the computational core of the system.

REFERENCES

- [1] D. N. Zmejcev, A. V. Klimov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, "Dataflow computing model as a paradigm of future mainstream of software development," *Informatics and Applications*, 2015, vol. 9, issue 4, pp. 29-36.
- [2] L. Yavits, A. Morad, R. Ginosar, "Computer architecture with associative processor replacing last level cache and simd accelerator", *IEEE Transactions on Computers*, Feb. 2015, vol. 64, issue 2, pp. 368-381. DOI=10.1109/TC.2013.220
- [3] M. Imani, D. Peroni, T. Rosing, "Nvalt: Nonvolatile approximate lookup table for GPU acceleration," *IEEE Embedded Syst. Lett.*, Mar. 2018, vol. 10, pp. 14-17.
- [4] A. D. Ivannikov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, D.N. Zmejcev, "Global Distributed Associative Environment - Evolution of Parallel Dataflow Computing System "Buran",", in *Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2018)*, Kazan, Russia, Sept 14 - 17, 2018, pp. 655-659.
- [5] A. D. Ivannikov, N. N. Levchenko, A. S. Okunev, A. L. Stempkovsky, D. N. Zmejcev, "Dataflow Computing Model – Perspectives, Advantages and Implementation," in *Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2017)*, Novi Sad, Serbia, Sept 29 - Oct 2, 2017, pp. 187-190.
- [6] M. Imani, D. Peroni, T. Rosing, "Nvalt: Nonvolatile approximate lookup table for GPU acceleration," *IEEE Embedded Syst. Lett.*, Mar. 2018, vol. 10, pp. 14-17.
- [7] Abbas Rahimi, Amirali Ghofrani, Kwang-Ting Cheng, Luca Benini, and Rajesh K. Gupta, "Approximate Associative Memristive Memory for Energy-Efficient GPUs," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Grenoble, France, March 9-13, 2015, pp. 1497-1502. DOI=10.7873/date.2015.0579
- [8] Mohsen Imani, "ReMAM: Low Energy Resistive Multi-Stage Associative Memory for Energy Efficient Computing", *17th International Symposium on Quality Electronic Design (ISQED)*, March 2016, pp. 101-106. DOI=10.1109/ISQED.2016.7479183
- [9] Rotem Ben Hur and Shahar Kvatinisky, "Memory Processing Unit for In-Memory Processing," in *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH 2016)*, Beijing, China, 18-20 July 2016, pp. 171-172. DOI=10.1145/2950067.2950086
- [10] Soroosh Khoram, Yue Zha, Jialiang Zhang, Jing Li, "Challenges and Opportunities: From Near-memory Computing to In-memory Computing", in *Proceeding of the 2017 ACM on International Symposium on Physical Design*, Portland, Oregon, USA, March 19-22, 2017, pp. 43-46. DOI=10.1145/3036669.3038242
- [11] Iacob Petrescu, "An FPGA based Associative Execution Unit," *The XXXVIII-International Scientific Symposium, Military Equipment and Technologies Research Agency*, Bucharest, 29-30 May, 2008.
- [12] Hasan Erdem Yantir, Ahmed M. Eltawil, Fadi J. Kurdahi, "A Two-Dimensional Associative Processor," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Sept. 2018, vol. 26, issue 9, pp. 1659-1670. DOI=10.1109/TVLSI.2018.2827262
- [13] H. E. Yantir, A. M. Eltawil, S. Niar, F. J. Kurdahi, "Power optimization techniques for associative processors," in *Journal of Systems Architecture*, October 2018, vol. 90, pp. 44-53. DOI=10.1016/j.sysarc.2018.08.006
- [14] L. Yavits, A. Morad, R. Ginosar, "Sparse matrix multiplication on an associative processor," *IEEE Transactions on Parallel and Distributed Systems*, 2015, vol. 26, issue 11, pp. 3175-3183. DOI=10.1109/TPDS.2014.2370055
- [15] Steven K. Esser, Paul A. Merolla, John V. Arthur, Andrew S. Cassidy, Rathinakumar Appuswamy, Alexander Andreopoulos, David J. Berg, Jeffrey L. McKinstry, Timothy Melano, Davis R. Barch, Carmelo di Nolfo, Pallab Datta, Arnon Amir, Brian Taba, Myron D. Flickner, and Dharmendra S. Modha, "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface," *Science* 08 Aug 2014, vol. 345, issue 6197, pp. 668-673. DOI=10.1126/science.1254642
- [16] Mingu Kang, Sujan K. Gonugondla, Ameya D. Patil, Naresh R. Shanbhag, "A Multi-functional In-memory Inference Processor using a Standard 6T SRAM Array," *IEEE Journal of Solid-State Circuits* Feb. 2018, vol. 53, issue 2, pp. 642-655. DOI=10.1109/JSSC.2017.2782087
- [17] I. Petrescu, D. Popescu, A. Petrescu, "A generic FPGA based associative processor," *17th INTERNATIONAL CONFERENCE ON*

- CONTROL SYSTEMS AND COMPUTER SCIENCE, 28 May 2009, Bucharest, Romania. URL: <http://rdsl.csit-sun.pub.ro/papers/p2009050002/CSCS17-C10.1.pdf>
- [18] L. Yavits, S. Kvatinsky, A. Morad, and R. Ginosar, "Resistive associative processor," *IEEE Comput. Archit. Lett.*, Jul. 2015, vol. 14, no. 2, pp. 148–151.
- [19] L. Yavits, A. Morad, R. Ginosar, "Associative Processor," supplementary material, 2014.
- [20] Hong Wang, Robert A. Walker, "Implementing a Scalable ASC Processor," in *Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS'03)*, Nice, France, 22-26 April 2003, p. 267a. DOI=10.1109/IPDPS.2003.1213482
- [21] Zbigniew Kokosinski, Wojciech Sikora, "An FPGA Implementation of a Multi-comparand Multi-search Associative Processor," in *Proc. 12th Int. Conf. FPL 2002, Lect. Notes in Comp. Sci.*, Springer-Verlag, 2002, vol. 2438, pp. 826-835.
- [22] Zbigniew Kokosinski, Bartłomiej Malus, "FPGA implementations of a parallel associative processor with multi-comparand multi-search operations," in *International Symposium on Parallel and Distributed Computing 2008*, 1-5 July 2008, Krakow, Poland, pp. 444-448. DOI=10.1109/ISPDC.2008.42
- [23] Avidan Akerib, "In-Place Associative Computing", electronic presentation. URL: <http://web.stanford.edu/class/ee380/Abstracts/171108-slides.pdf>
- [24] D. E. Yakhontov, N. N. Levchenko, A. S. Okunev, "Principles of work of the special operations unit of the associative memory module of parallel dataflow computing system," in *Proc. Supercomputer technologies: design, programming, application (SCT-2010)*, Taganrog, Moscow, 2010, vol. 1, pp. 166—170. (In Russian).
- [25] Trong Tu Bui, Tadashi Shibata, "Low-Power Analog Associative Processors Employing Resonance-Type Current-Voltage Characteristics," *Solid State Circuits Technologies*, Book edited by: Jacobus W. Swart, ISBN 978-953-307-045-2, January 2010, INTECH, Croatia, pp. 462.
- [26] Trong Tu Bui, Tadashi Shibata, "A multi-core/multi-chip scalable architecture of associative processors employing bell-shaped analog matching cells," in *Proc. 9th International Conference on Solid-State and Integrated-Circuit Technology*, Beijing, China, 20-23 Oct. 2008, DOI=10.1109/ICSICT.2008.4734933
- [27] M. Imani, "MASC: Ultra-low energy multiple-access single-charge TCAM for approximate computing," *IEEE/ACM Design Automation & Test in Europe Conference & Exhibition (DATE)*, 2016, pp. 373-378.
- [28] N. N. Levchenko, A. S. Okunev, D. E. Yakhontov, D. N. Zmejev, "Decreasing the power consumption of content-addressable memory in the dataflow parallel computing system," in *Proceeding of IEEE EAST-WEST DESIGN & TEST SYMPOSIUM 2012*, Kharkov, Ukraine, September 14-17, 2012, pp. 122-125. DOI=10.1109/ewdts.2013.6673191
- [29] M. Imani and T. Rosing, "CAP: Configurable resistive associative processor for near-data computing," in *Proc. 18th Int. Symp. Quality Electron. Design (ISQED)*, Mar. 2017, pp. 346–352.
- [30] Mohsen Imani, A. Rahimi and Tajana Rosing, "Resistive configurable associative memory for approximate computing," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 14-18 March 2016, Dresden, Germany, pp. 1327-1332, URL: http://cseweb.ucsd.edu/~abrahimipubs/DATE16_ReCAM.pdf
- [31] Mohsen Imani, Shruti Patil, Tajana Šimunić Rosing, "Approximate Computing Using Multiple-Access Single-Charge Associative Memory," *Emerging Topics in Computing IEEE Transactions on*, 2018, vol. 6, no. 3, pp. 305-316.

SWIELD: An In Situ Approach for Adaptive Low Power and Error-Resilient Operation

Mitko Veleski¹, Rolf Kraemer^{1,2} and Milos Krstic^{2,3}

¹BTU Cottbus-Senftenberg, Cottbus, Germany

²IHP - Leibniz-Institut für innovative Mikroelektronik, Frankfurt (Oder), Germany

³University of Potsdam, Potsdam, Germany

E-mail: {veleski, kraemer, krstic}@ihp-microelectronics.com

Abstract—With the rapid CMOS technology downscaling, two requirements for digital integrated circuits (ICs) design are becoming fundamental: low power consumption and error resilience. Many proposed solutions are effective either in reducing the power consumption or improving the error resilience, but a single approach that addresses both aspects seems to be missing. The reason behind this might be the often opposing relationship between the two requirements. However, employing adaptive techniques that adjust the systems' behaviour according to the current needs might be promising. We propose SWIELD - a programmable in-situ monitor (ISM), i.e. a monitor placed inside the actual circuit, able to operate in three modes: Normal, AVS and TMR. The architecture and the operation modes of the SWIELD are described in detail. A specially designed SWIELD operation control unit is also introduced. The proposed ISM design is synthesized using the IHP 130nm technology library. Furthermore, a time-based power consumption analysis is performed. The obtained values are then compared to a similar ISM design. At the expense of small area and complexity overhead, the SWIELD design shows satisfactory results in terms of power saving and adaptivity.

I. INTRODUCTION

Error resilience of the digital integrated circuits (ICs) has been traditionally a crucial requirement in specific electronic systems applications such as space, avionics or automotive industries. With the aggressive CMOS technology downscaling, the error resilience becomes increasingly important also for less critical electronic systems-based applications. Furthermore, the enormous number of integrated components on a single chip introduces excessive power consumption. Therefore, an adequate handling with these challenges is of utmost importance.

The industry and research communities have been actively working on optimizing power consumption and enhancing error resilience. However, the proposed techniques mainly address these aspects separately, that is, the focus is put either on the error resilience or on the power consumption. An intriguing trade-off between the two requirements aggravates the possibility to meet them both simultaneously. Most of the approaches that improve the error-resilience (particularly to externally induced soft errors or to aging induced permanent errors) use some form of redundancy, but this leads to increased power consumption. On the other hand, utilizing techniques like voltage scaling in order to reduce the power

consumption has shown to affect the error resilience negatively in two aspects. First, reducing the supply voltage slows down the switching speed of the transistors which results in timing errors. Failing to meet the timing constraint is unacceptable and may cause catastrophic consequences. Second, if the voltage is scaled, the transistors have lower noise margins that make the circuit more prone to electromagnetic interference and noise, while the conductivity of the metal conductors is lowered [1], [2]. Furthermore, the soft-error rate (SER) increases exponentially with the downscaling of the supply voltage [3], [4]. Thus, a synergistic approach that would encompass both reliability and power consumption is necessary.

Additionally, static and dynamic variations referred to as PVTa (Process, Voltage, Temperature and Aging) variations are more pronounced as the CMOS technology continues to scale down. PVTa variations also impose a threat to the digital ICs resilience manifested mainly as increased path delay, which can in turn lead to timing errors. The traditional approach to avoid timing errors is known as guard-banding, i.e. insertion of an additional safety voltage margin, to guarantee correct operation under worst-case conditions. Since the worst-case scenario occurs rather rarely, guard-banding leads to unnecessary high power consumption and significant energy waste. It is, therefore, preferable to optimize the power consumption by adjusting the supply voltage according to the current operating/environmental conditions. Moreover, employing an adaptive technique able to accommodate to the application requirements could possibly maintain a balanced level of resilience and power consumption in the system.

Unlike conventional low-power approaches, Adaptive Voltage Scaling (AVS) is able to save a notable amount of energy as well as to prevent excessive voltage reduction. It is an advanced low-power closed-loop technique that enables tuning of the supply voltage based on a feedback information from a special circuitry that constantly observes the system operation [5], [6]. The special circuits responsible for observing the system operation by measuring the path delay could be implemented as replica (canary) circuits or as *in situ* monitors. The former [7], [8] use ring-oscillators or some other form of delay lines to mimic the most critical path in the system. Adjusting of the supply voltage is, thus, performed by measuring the speed of the replica path instead of the real critical

path. Such approach is, however, only suitable for tracking the global variations. Local variations affect the timings of the replica path and the real critical path differently. This inconsistency would eventually lead to inappropriate supply voltage management. Unlike the replica circuits, the in situ monitors (ISMs) are placed inside the actual circuit which enables them to keep track of the local variations. This is usually achieved by replacing the timing critical flip-flops with ISMs.

In this paper, we propose a new approach to the design of in situ monitors referred to as SWIELD. Depending on the current requirements, the proposed ISM could be used either as timing error predictive monitor (necessary for correct AVS implementation) or as TMR flip-flop. If, however, there is no need of such enhanced functionalities, it can also behave as a regular flip-flop. The key idea behind SWIELD is the programmability feature, that is, the system can easily switch between its operation modes in order to optimize the power consumption or to enable error protection. SWIELD provides three operation modes: Normal, AVS and TMR. The switching between the operation modes is achieved by writing a register in a specially designed control unit which serves both as an AVS controller and as a SWIELD operation manager. The proposed design could be especially useful in complex dependable systems which require low power consumption and error resilience. Furthermore, SWIELD is suitable for real-time applications because it predicts the timing errors and hence, does not require error recovery mechanisms which introduce performance penalties.

The rest of the paper is structured as follows. Section II gives an overview of the related work in this area. In Section III the architecture and the operation modes of the proposed SWIELD ISM are described. We present the Extended Voltage Scaling Control Unit responsible for managing the SWIELD and AVS operation in Section IV. The practical results and discussions are given in Section V. Finally, we conclude the paper and outline the directions for future work in Section VI.

II. RELATED WORK

ISMs can be designed either to detect timing/soft-errors or to predict timing errors. The error detection approach usually requires less complexity, area and power. However, it introduces performance penalties as some error recovery mechanisms have to be employed to correct the detected errors. Often, such performance degradation is not acceptable. On the other hand, the error prediction approach requires more hardware per monitor, but is able to warn the system before the occurrence of a timing error, and therefore performance penalties are avoided. After the reception of the warning signal, the system through the AVS controller could take some action, e.g. increasing the supply voltage, to prevent the timing error from occurring.

One of the first and most significant works which rely on ISMs is Razor [9]. It consists of replacing the flip-flops which reside at the end of the critical paths with so-called Razor flip-flops. A Razor flip-flop is actually an ISM which

contains the regular flip-flop augmented with a shadow latch connected in parallel. The outputs of the regular flip-flop and the shadow latch are compared by a comparator (eg. XOR gate). While the regular flip-flop samples the data input signal on the active clock edge, the latch is transparent during the whole clock duty cycle. Therefore, if the input data transition arrives early enough to meet the flip-flop setup time, both the regular flip-flop and the shadow latch hold the same value which indicates correct operation. However, if the input data transition arrived late to meet the regular flip-flop setup time, the shadow latch still captures the correct data value. Hence, the compared output values differ and as a result, a timing error is signalled by the comparator. According to the timing error rate, the system adjusts the supply voltage level.

The Razor approach made major breakthrough related to dealing with both global and local variations as well as optimizing power consumption using ISM-based AVS - up to 42% energy savings is reported by the authors. However, it suffers from several drawbacks: as it is an error detection-based technique, it requires error recovery mechanisms which introduce area overhead and performance degradation. Furthermore, during longer periods without detected timing errors, the supply voltage could be overscaled which might lead to increased timing error rate later on. Additional problems related to Razor are metastability and short-path constraint, i.e. mistaking a fast but correct transition for a timing error. Due to these disadvantages, Razor is not eligible for real-time applications.

The potential of ISMs was recognized quite promptly and substantial amount of research work has been conducted to improve the design and to overcome the limitations of Razor. The authors in [10] introduce RazorII - a simplified ISM design which overcomes the short-path problem and focuses only on error detection, while the error recovery is performed on the architecture level. RazorII is capable of detecting both timing errors and Single Event Upsets (SEUs) and reports energy savings of approximately 33%.

In [11], the authors propose two ISM-based error-detecting circuits: dynamic transition detector with time-borrowing (TDTB) and double-sampling static design with a time-borrowing (DSTB) datapath latches. The main advantage of this work in comparison to Razor is the drastic alleviation of the metastability problem. Instruction replay at lower clock frequency is utilized as error recovery mechanism. Total power reduction of up to 37% is reported.

An ISM design referred to as SETTOFF - Soft-Error and Timing error Tolerant Flip-Flop is introduced in [12]. At the expense of increased complexity and power consumption compared to Razor and RazorII, SETTOFF is capable of detecting Single Event Transients (SETs), timing errors and SEUs. Additionally, an on-the-fly SEU recovery is enabled by inverting the main flip-flop state immediately after the error detection. However, power optimization and savings are not provided.

Another ISM proposal able to detect and recover from both soft and timing errors is described in [13]. The three flip-

flops that form the ISM use clocks with same frequency, but different phases. Such approach is quite unusual and might be problematic. However, it is reported that the proposed design might improve the system performance by overclocking. Here, optimization of the power consumption is not considered at all.

A timing error-predictive ISM for AVS implementation referred to as Pre-Error Flip-Flop is proposed in [14]. Unlike the other error detection, Razor-like designs, the Pre-Error approach is able to detect late, but non-erroneous data transitions (pre-errors). Thus, it can predict the timing error before it happens. Similarly to Razor, however, the Pre-Error Flip-Flop besides the regular flip-flop, contains another flip-flop in parallel (Figure 1). Since the error warnings are issued before the occurrence of an actual error, no logic for recovery is needed. Hence, additional hardware overhead and performance penalties introduced by error recovery computations are avoided. This makes the Pre-Error ISM suitable for real-time applications.

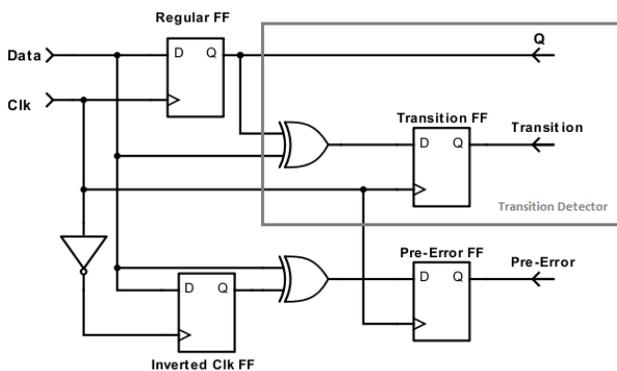


Fig. 1: Schematic diagram of the Pre-Error ISM

A data transition is considered a pre-error if it occurs during a period called pre-error detection window. This is a time interval before the active edge of the clock. It is crucial to pick a length for the detection window as accurate and as robust as possible. To define the detection window length, the authors exploit the clock duty cycle which means that the detection window starts with the falling edge. Changing the duty cycle is, however, possible only if the falling edge does not trigger logic events in the system. Another error-predictive ISM approach with ability to tune the detection window length is introduced in [15].

Voltage scaling is performed based on the pre-error rate which is an indicator for the speed of the circuit. When the error rate is zero or near-zero, the AVS controller can reduce the supply voltage. On the other hand, high error rate requires increasing the voltage level. If the error rate keeps some medium value, the voltage can be held constant. However, since voltage tuning is performed while the system is on-line, idle or low-activity periods would result in excessive voltage decrease. This would most likely lead to increased timing error rate later when the system activity intensifies. To prevent the

voltage overscaling problem from occurring (specific to Razor-like approaches), a so called transition detector observes the activity of the circuit (Figure 1). The transition detector is also part of the Pre-Error Flip-Flop and plays an important role in the voltage scaling process.

Tested on multiplier and Discrete Cosine Transform (DCT) circuits, the Pre-Error ISM design is reported to yield up to 36% power savings at the expense of minimal overhead introduced by the additional hardware necessary to build the ISM and the AVS controller. Furthermore, the performance is preserved as the timing errors are avoided. However, the Pre-Error ISM, as proposed in [14] is unable to deal with soft errors.

As can be noted, the previously published works focus either on optimizing power consumption or on error resilience. Our ISM design is capable of saving power and providing error resilience by simply switching between the adequate operation modes depending on the application and environmental requirements. The baseline for this work is the concept proposed in [14]. We utilize the redundant hardware within the original ISM and improve its functionality by making it programmable. The newly designed ISM, metaphorically speaking, cuts the additional voltage safety margins like a SWORD and protects against errors like a SHIELD and thus, the inspiration to name it SWIELD.

III. SWIELD ARCHITECTURE AND OPERATION MODES

As illustrated in Figure 2, the programmability feature of the SWIELD ISM is provided by introducing two special input signals: TMR and Gated Clk. A dedicated component called Extended Voltage Scaling Control Unit (EVSCU) described in Section IV is responsible for driving these signals.

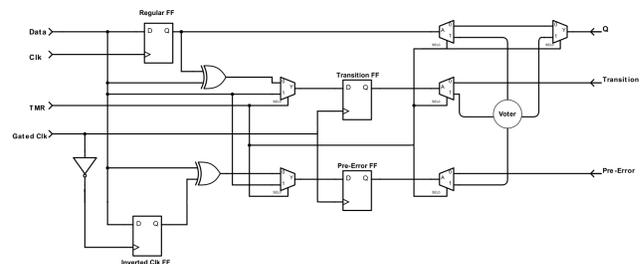


Fig. 2: Schematic diagram of the SWIELD ISM

The SWIELD Data input signal is the adequate input signal to the Regular flip-flop. The output signal Q could reflect either the output from the Regular flip-flop or the output from the voter depending on the current operation mode. The three additional flip-flops (Inverted Clk FF, Transition FF and Pre-Error FF) are necessary to implement the basic Pre-Error ISM structure (Figure 1). The Inverted Clk FF is crucial for providing the pre-error detection window, while the Transition FF and Pre-Error FF are responsible for sampling the Data input activity and transitions within the detection window respectively. Figure 3 illustrates how the Transition and Pre-

Error outputs are generated when the SWIELD ISM is used to observe the system operation.

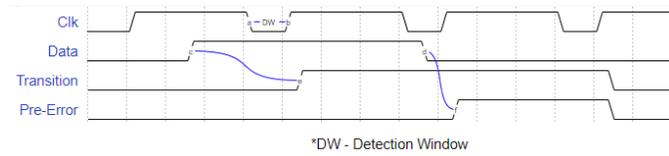


Fig. 3: Generation of Transition and Pre-Error outputs

The Transition and Pre-Error outputs from every SWIELD ISM in the system should be orred together and then connected to the EVSCU. Thereby, the overall Transition and Pre-Error rates would be provided. Note that the programmability feature of the SWIELD ISM comes at a price of some additional circuitry (multiplexers, demultiplexers and a voter to enable TMR). The three modes of operation are elaborated in the next subsections.

A. Normal mode of operation

When the TMR input is logical '0' and the clock gating is enabled, the SWIELD ISM operates in Normal mode, i.e. only the Regular FF is active. This mode is useful in scenarios where the voltage scaling is not performed, i.e., the supply voltage is held constant. Hence, substantial amount of energy is saved by clock-gating the additional flip-flops within the SWIELD ISM. The Q output of the ISM reflects the output of the Regular FF, while the Transition and Pre-Error outputs are logical '0' in this mode.

B. AVS mode of operation

This mode of operation is similar as described in [14]. Thus, every SWIELD ISM observes the critical path delay and sends feedback to the EVSCU (the AVS controller) through Transition and Pre-Error outputs (Figure 3). Based on this information, the EVSCU instructs the voltage regulator to adequately adjust the supply voltage. Here again the Q output from the SWIELD ISM is equal to the output of the Regular FF. The TMR input is logical '0' and the clock gating is disabled in this scenario.

C. TMR mode of operation

If the TMR input is set to logical '1', the Regular FF together with the Transition FF and Pre-Error FF form a Triple Modular Redundancy (TMR) structure. In order to activate this mode of operation, two pre-conditions are required: first, the EVSCU has to disable the voltage scaling, that is, to fix the supply voltage to a constant level and second, the clock gating must be disabled. Thus, the TMR mode of operation enables protection against soft errors. It is useful to switch to TMR mode when the system is (expecting to be) exposed to higher soft error rates. The Q output of the SWIELD ISM in this mode is driven by the voter output, while the Transition and Pre-Error outputs are logical '0'.

IV. EXTENDED VOLTAGE SCALING CONTROL UNIT

The SWIELD ISM could be integrated into any general-purpose electronic system. Of particular interest is integration into complex, (multi)processor-based System-On-Chip (SoC). Within such design, a simple and flexible dedicated control unit called Extended Voltage Scaling Control Unit (EVSCU) and a voltage regulator must be connected in order to enable implementation of AVS. The EVSCU is crucial for providing high level of system adaptivity and has two functions: it serves as a driver of the voltage regulator and is also responsible for managing the operation modes of the SWIELD ISMs. A block-diagram of the EVSCU is shown in Figure 4.

Inputs to the EVSCU are the orred Transition and Pre-Error signals. As shown in Figure 4, the EVSCU contains a register set which can be written or read-out via the system internal data bus. The Transition and Pre-Error inputs together with the values stored in the register set provide the necessary information to drive the Clock Gating and Voltage Scaling Logic blocks.

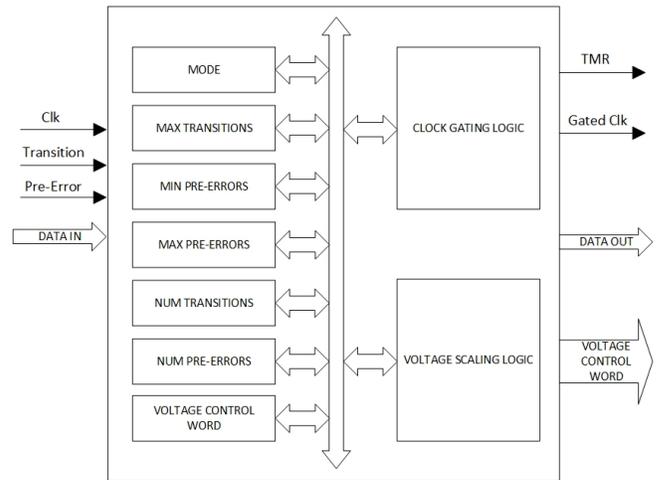


Fig. 4: Block-diagram of the Extended Voltage Scaling Control Unit (EVSCU)

The operation mode of the SWIELD ISMs is set by writing the MODE register. Based on the current operation mode, the TMR and Gated Clk outputs (which are actually inputs to every SWIELD ISM) are adequately generated. By writing the MAX TRANSITIONS, MIN PRE-ERRORS and MAX PRE-ERRORS registers, the parameters required by the Voltage Scaling Logic block are defined. The NUM TRANSITIONS and NUM PRE-ERRORS are read-only registers that contain the current numbers of transitions and pre-errors respectively. Thus, when the system operates in AVS mode, the Transition and Pre-Error inputs are observed during a time period of MAX TRANSITIONS. Then, the Voltage Scaling Logic compares the current number of pre-errors to the values in the MIN PRE-ERRORS and MAX PRE-ERRORS registers. If the current number of pre-errors is lower than MIN PRE-ERRORS, then the Voltage Scaling Logic instructs the voltage regulator to decrease the supply voltage for one step by gen-

erating adequate value for the VOLTAGE CONTROL WORD output. If it is higher than MAX PRE-ERRORS, the supply voltage is increased for one step in a similar way. The supply voltage level does not change if the current number of pre-errors is between the values stored in MIN PRE-ERRORS and MAX PRE-ERRORS registers. Finally, when the voltage scaling is not performed, i.e. the system operates in one of the two remaining operation modes, the desired level of supply voltage can be set simply by writing the VOLTAGE CONTROL WORD register.

V. RESULTS AND DISCUSSION

To investigate the benefits of the programmable SWIELD ISM, first we synthesize the designs and compare the reported data for both Pre-Error and SWIELD ISMs. Then, we run an exhaustive post-synthesis simulation in order to analyze the time-based power consumption of both designs. Finally, we compare the obtained results. We chose the Pre-Error ISM for comparison because it has similar structure and features to the SWIELD ISM. Furthermore, the SWIELD ISM shares similar behaviour with the Pre-Error ISM more than any other ISM design described in Section II. The synthesis is performed using the IHP 130nm technology library [16].

	Pre-Error ISM	SWIELD ISM
Combinational Cell Count	11	16
Sequential Cell Count	4	4
Combinational Area (μm^2)	85.05	147.42
Noncombinational Area (μm^2)	120.96	120.96
Overall Design Area (μm^2)	214.1	282.82
Dynamic Power Dissipation (μW)	3.8	4.1

TABLE I. Synthesis reports comparison of Pre-Error and SWIELD ISMs (Technology: IHP 130nm; Nominal Supply Voltage: 1.2V; Frequency: 50MHz).

Table I shows the data obtained after synthesis of the designs for Pre-Error and SWIELD ISMs. Under the same conditions, the SWIELD ISM occupies 32% more area and dissipates almost 8% more dynamic power than the Pre-Error ISM. Such outcome is expected as the SWIELD ISM contains additional logic that provides the programmability feature. Note that the dynamic power dissipation is statically estimated value without considering the switching activities of the designs.

Now, let us take a look at the time-based power consumption of the Pre-Error and SWIELD ISMs. In order to obtain credible results, we run a post-synthesis simulation based on an exhaustive testbench (the same testbench applies for both designs). Furthermore, the simulation takes into account the switching activities of the ISMs defined by the stimuli in the testbench. The time-based power consumption is calculated by power analysis tool which takes a switching activity file as an input.

Based on the data shown in Table II, one can conclude that the Pre-Error ISM is more power-efficient when the operation mode is set to AVS (the only operation mode available to the

		Operation Mode			
		AVS	TMR	Normal	All three combined
Power Consumption (μW)	Pre-Error ISM	6.54	N.A.	N.A.	N.A.
	SWIELD ISM	8.33	7.08	2.26	5.88

TABLE II. Time-based power consumption comparison of Pre-Error and SWIELD ISMs (Technology: IHP 130nm; Nominal Supply Voltage: 1.2V; Frequency: 50MHz).

Pre-Error ISM). This is also expected due to the additional logic contained in the SWIELD ISM. On the other hand, when the operation mode of the SWIELD ISM is switched to Normal, the overall power consumption is reduced almost four times (compared to the AVS mode). However, the most intriguing is the case when all the three operation modes are switched in one simulation run (all three get an equal portion of the simulation time). The overall SWIELD ISM power consumption is then 42% and 20% lower than AVS- and TMR-only cases respectively. Moreover, it is 11% lower than the Pre-Error ISM running the same testbench in its only available operation mode. This result confirms the advantage of employing such adaptive design - the operation mode is switched according to the current application needs and environmental conditions, while the power consumption is still optimized. To the best of our knowledge, none of the previously published related works is able to achieve this.

VI. CONCLUSION AND FUTURE WORK

In this paper we presented SWIELD - a programmable ISM able to operate in three modes: Normal, AVS and TMR. An architecture of a simple and highly flexible control unit which manipulates the SWIELD operation is also proposed. The results presented in Section V show substantial potential for power saving having in mind the adaptivity and flexibility provided by the different operation modes. One of the main challenges that need to be addressed in the future is to investigate the relationship between the timing-critical and soft error-critical flip-flops. This is crucial when integrating the SWIELD ISM into a complex design. The number of the replaced flip-flops with ISMs has to be minimal, yet effective. Evaluation of the resilience to soft errors as well as integration of the SWIELD into complex (multi)processor SoC is also left for a future work.

REFERENCES

- [1] Y. S. Dhillon, A. U. Diril, and A. Chatterjee, "Soft-error tolerance analysis and optimization of nanometer circuits," in *Design, Automation, and Test in Europe*, pp. 389–400, Springer, 2008.
- [2] H. Wang, S. V. Rodriguez, C. Dirik, and B. Jacob, "Electromagnetic interference and digital circuits: An initial study of clock networks," *Electromagnetics*, vol. 26, no. 1, pp. 73–86, 2006.
- [3] D. Zhu, R. Melhem, and D. Mossé, "The effects of energy management on reliability in real-time embedded systems," in *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*, pp. 35–40, IEEE, 2004.

- [4] A. Ejlali, M. T. Schmitz, B. M. Al-Hashimi, S. G. Miremadi, and P. Rosinger, "Energy efficient seu-tolerance in dvs-enabled real-time systems through information redundancy," in *Low Power Electronics and Design, 2005. ISLPED'05. Proceedings of the 2005 International Symposium on*, pp. 281–286, IEEE, 2005.
- [5] D. Flynn, R. Aitken, A. Gibbons, and K. Shi, *Low power methodology manual: for system-on-chip design*. Springer Science & Business Media, 2007.
- [6] B. Amrutur, N. Mehta, S. Dwivedi, and A. Gupte, "Adaptative techniques to reduce power in digital circuits," *Journal of Low Power Electronics and Applications*, vol. 1, no. 2, pp. 261–276, 2011.
- [7] T. D. Burd, T. A. Pering, A. J. Stratakos, and R. W. Brodersen, "A dynamic voltage scaled microprocessor system," *IEEE Journal of solid-state circuits*, vol. 35, no. 11, pp. 1571–1580, 2000.
- [8] M. Elgebaly and M. Sachdev, "Efficient adaptive voltage scaling system through on-chip critical path emulation," in *Proceedings of the 2004 International Symposium on Low Power Electronics and Design (IEEE Cat. No. 04TH8758)*, pp. 375–380, IEEE, 2004.
- [9] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, *et al.*, "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture*, p. 7, IEEE Computer Society, 2003.
- [10] S. Das, C. Tokunaga, S. Pant, W.-H. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, and D. T. Blaauw, "RazorII: In situ error detection and correction for pvt and ser tolerance," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, 2008.
- [11] K. A. Bowman, J. W. Tschanz, N. S. Kim, J. C. Lee, C. B. Wilkerson, S.-L. L. Lu, T. Karnik, and V. K. De, "Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 49–63, 2009.
- [12] Y. Lin and M. Zwolinski, "Settoff: A fault tolerant flip-flop for building cost-efficient reliable systems," in *2012 IEEE 18th International On-Line Testing Symposium (IOLTS)*, pp. 7–12, IEEE, 2012.
- [13] N. D. P. Avirneni and A. Somani, "Low overhead soft error mitigation techniques for high-performance and aggressive designs," *IEEE Transactions on Computers*, vol. 61, no. 4, pp. 488–501, 2011.
- [14] M. Wirnshofer, *Variation-aware adaptive voltage scaling for digital CMOS circuits*. Springer, 2013.
- [15] W. Shan, L. Shi, and J. Yang, "In-situtiming monitor-based adaptive voltage scaling system for wide-voltage-range applications," *IEEE Access*, vol. 5, pp. 15831–15838, 2017.
- [16] <https://www.ihp-microelectronics.com>.

Automating of Human Resources Management using Genetic Algorithms

Agata V. Markevich,
PhD Student, of Department
«Control and Information Protection»,
Federal State Institution of Higher Education
«Russian Transport University» (MIIT)
Moscow, Russia
vlasjuk.a@mail.ru

Valentina G. Sidorenko,
Professor of Department
«Control and Information Protection»,
Federal State Institution of Higher Education
«Russian Transport University» and
Department of Modeling and Business Process
Optimization Higher School of Economics
Moscow, Russia
valenfalk@mail.ru

Abstract — Forecasting to improve resource efficiency is an important element for decision making. Based on predictive estimate we can form a performance management system and optimize the financial costs of team payroll.

The article presents the objective setting and solution of finding the optimal allocation of project team resources according to various criteria using genetic algorithm. Also we will set proposals for adapting created model to solving other human resource management tasks in the enterprise.

Keywords—automated information-management systems, project management, resource allocation

I. INTRODUCTION

Discussing total and universal automation, we should understand that it develops gradually: (1) pointwise (solving separate tasks by digital means), (2) piecewise (combining tasks into blocks); (3) process (when the entire process is being automated); (4) integrally (no “manual” labor of employees/administrators is left between different processes, they are replaced by integration between processes).

Numerous products that enable accounting employees’ time are now available at the market. They imply fixing the time spent in the office, including integration with facial recognition systems, run screen screening, analysis of links visited in the Web. At the same time, the analysis of the working activity, namely, its rate and quality, does not get the sufficient attention.

Numerous work time logging applications are available, but an understanding of whether employees really use their working time with the maximal efficiently, or whether it is composed of alternating downtime and re-processing, is not always present [1].

The key operational task of personnel management is the staff time distribution and accounting. This task solution is used in the piecework remuneration calculation as the basis for the team standardization, i.e. it directly affects the organization cost reduction. When the resource allocation variability occurs, it is usual to solve such task by enumerative technique or other mathematical methods of seeking an optimal solution that are included in specialized applications [2]. The task entry conditions are formal resource allocation rules (shift work or fixed working hours, availability, or work task schedule) and various

restrictions (on project performance time, allowable expenses, number of team members).

At the same time, the possibility of operative introducing of additional restrictions and the ability of total cost quick assessment are important for the application use at the enterprise.

The entire line of assessment applications and methods solving the task set is available nowadays. Large enterprises use comprehensive solutions that integrate into the common IT landscape. Comprehensive solutions are provided by SAP, Oracle and other companies. However, small and medium business is not ready to overpay for integrated solutions that don’t always meet all the specific customer’s requirements and are aimed at standard queries and processes. Most local solutions are based on the use of Excel spreadsheets. In this case, the resource allocation is done with macros. Each of these applications has both advantages and disadvantages, mainly due to the limited flexibility of setting the input data and calculation principles [3]. The particular interest is raised by AFM program: Scheduler 1/11 application, designed for non-standard work schedules building and optimization, that allows setting various operation modes, account the work of employees of different specializations, leave schedule, and other parameters. However, this application cannot still be considered universal [4].

It is noteworthy many enterprises don’t use automation equipment or use applications developed independently for local tasks. Therefore, the task of seeking an optimal distribution of project teams’ resources remains topical. This article is dedicated to the stated task solution.

II. EASE OF USE ANALYSIS OF THE EXPERIENCE OF CAIMSS IMPLEMENTATION AT AN INDUSTRIAL ENTERPRISE

Computer-aided information management systems (CAIMSS) represent a part of modern industrial enterprise automation.

CAIMSS are widely used at modern industrial enterprises, along with numerous of automation tools. Articles [2, 5] consider the information society issues and determine the place of technical solutions in the operational and strategic management process.

The key requirement to CAIMSS being implemented is confidence in data storage reliability, the correctness of built-

in methodology and processes, and the user interface convenience. When the competition is high, industrial enterprises are forced not only to use modern software, but also to implement CAIMs in maximally short terms.

Let us consider the implementation process. Depending on the composition and number of management processes planned to be automated, and, consequently, the composition and number of CAIMS modules are planned to be implemented, and the enterprise features, primarily determined not by its size and scope of activity, but by the complexity of management processes, this process happens in different ways. In personnel management related issues, enterprises are increasingly choosing SAP SuccessFactors cloud solutions characterized by relatively low cost of implementation and support of final solutions, and shorter project implementation terms compared to classic SAP products. Figure 1 represents a typical deployment diagram for CAIMS SAP SuccessFactors [6].

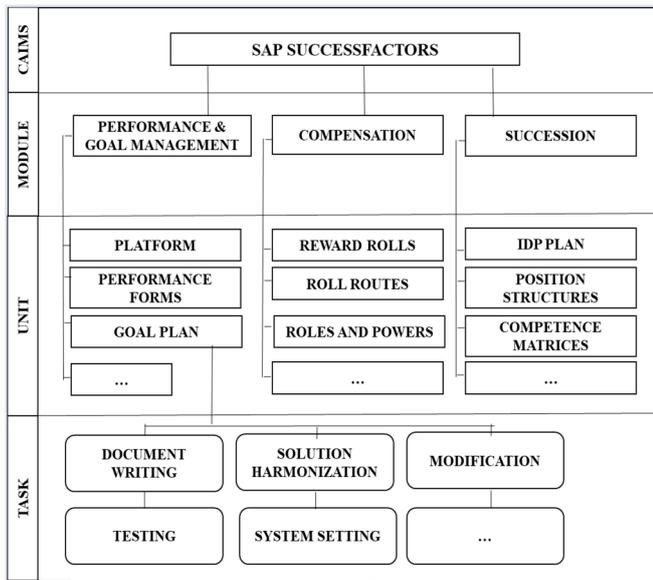


Fig. 1. Common project framework

CAIMS consists of a set of modules (goal setting and achievement assessment, remuneration management, talent management, analytics and strategic talent management, external career portal, recruitment and adaptation, HR training and development, HR administration), which can be implemented individually or as an integrated system. According to the implementation method, modules are divided into units (form setting, form route setting, users training). In order to implement separate units, it is required to perform a line of tasks (to coordinate, configure, test, etc.). Due to modularity, finiteness of possible setting combinations and the solution scaling simplicity, the implementation process is standardized and does not significantly differ from implementation to implementation. Figure 2 represents a diagram of the goal setting and achievement assessment module implementation (one of typical implementations), the analysis of the implementation experience of which this article is based on.

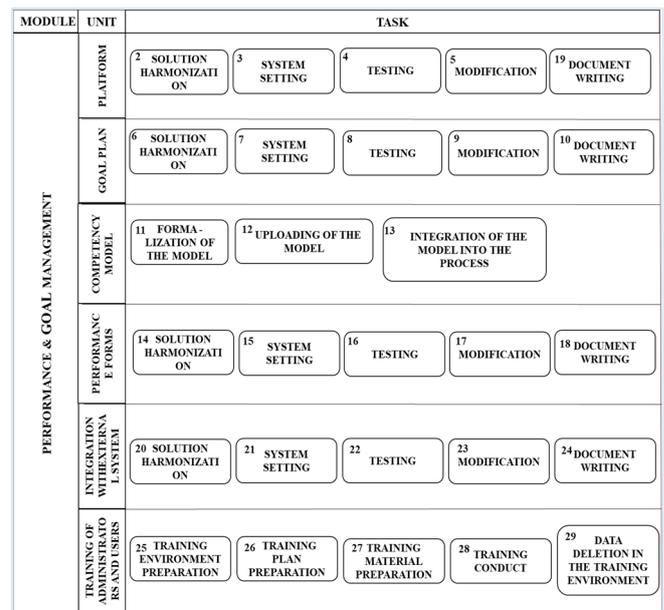


Fig. 2. Example of the process of goal setting and achievement assessment module implementation

In this case, the relatively fast implementation of processes with a pre-formed project solution based on the company formalized processes is described. If the customer has the wish and ability to optimize the processes, then, on the one hand, the project is complicated by a preliminary analysis of current processes and their optimization for automation, and on the other hand, this simplifies the subsequent implementation due to the product specifics.

Table 1 represents the characteristics of companies that introduced the goal setting and achievement assessment module for the period of 2015-2017, and the characteristics of the process being implemented. The author of the article participated in projects as a key performer, an architect or a project manager.

TABLE I. CHARACTERISTICS OF THE IMPLEMENTED PROJECTS OF THE GOAL SETTING AND ACHIEVEMENT ASSESSMENT PROCESS AUTOMATION

	Implementation №				
	1	2	3	4	5
1. Characteristics of companies					
Branch	<i>banking</i>	<i>chemical industry</i>	<i>retail</i>	<i>metallurgy</i>	<i>metallurgy</i>
The number of the test group users	up to 200	up to 200	up to 200	20 000	up to 200
2. Process characteristics					
Availability of a process of setting goals and evaluating achievements in the company at the beginning of the project	no	yes	no	yes	no
Preliminary process optimization	yes	no	yes	no	yes
The level of complexity of the process built into the system*	1	4	2	5	3
Number of additionally developed elements/extensions	0	3	1	12	0

3. Project characteristics					
Planned implementation term (month)	4	5	4	6	3
Planned number of consultants in the project team	3	2	2	3	3,5
Planned labor intensity (months of consultants' work)	16	15	12	36	15
Actual implementation term (month)	6	11	4	12	3,5
Actual number of consultants in the project team	4	3	3	6	5
Actual labor intensity (months of consultants' work)	24	33	12	72	17,5
Number of project team members from the customer's side	2	2	1	6	7

*expert assessment built on:

- Number of subprocesses (goal setting, monitoring, etc.);
- Number of assessment periods (per year);
- Goal setting frequency (per year);
- Number of unique elements in the system (efficiency forms, goal plans, etc.);
- Number of various roles in the process.

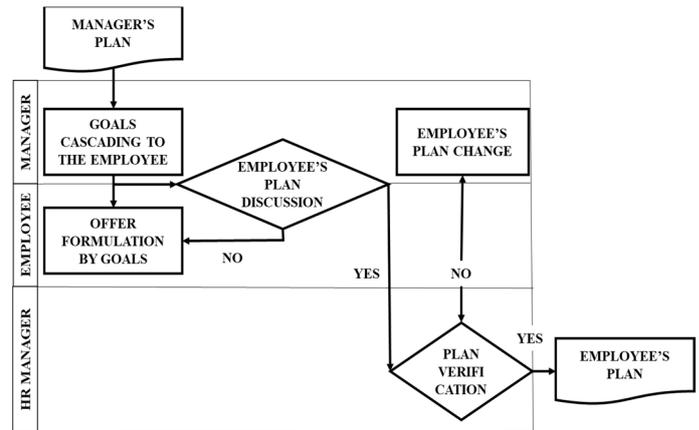
Main conclusions to Table 1:

- If the company has an established methodology and formalized processes for goal setting and achievement assessment, the customer agrees for their optimization reluctantly;
- The system is well-scalable, so the number of the test group users does not affect the process level of complexity directly;
- With the simultaneous implementation of the method of goal setting and achievement assessment, assessment of relevant processes and a process automation system, it is possible to significantly reduce the risks related the project terms prolongation due to the need of developing additional elements and enhancing the project team;
- The number of project team members from the customer's side does not affect the project timing directly;
- Risks related the project terms prolongation (and, therefore, cost increase) can be reduced by improving the quality of pre-project analysis aimed at determining the number of additional elements enhancing the standard CAIMS functionality, and by optimizing the expenses for consultants.

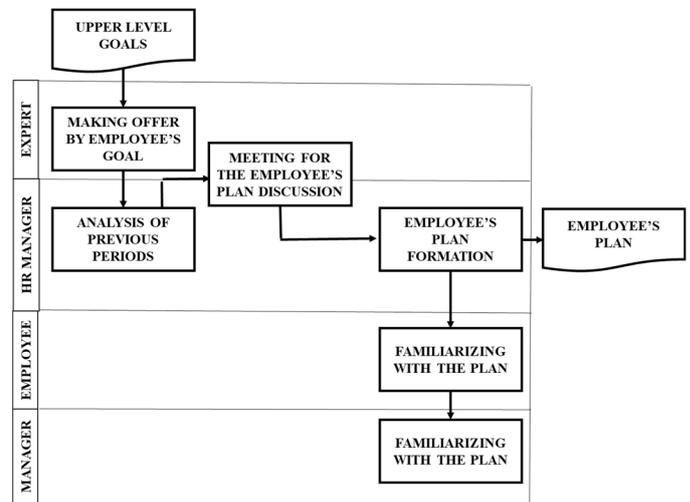
As a rule, the goal setting and achievement assessment process includes the following subprocesses: goal setting, goal implementation monitoring, achievement assessment, and calibrating of results at the HR committee.

The goal setting process is shown on Figure 3:

- Typical for global practices in-built into SAP SuccessFactors (a);
Requiring the creation of additional elements (extensions) (b).



(a)



(b)

Fig. 3. (a) Fragment of the goal-setting process automated within the of the achievement setting and assessment module framework without outside extensions; (b) Fragment of the goal-setting process automated within the of the achievement setting and assessment module framework with outside extensions

Despite the product modularity and adaptability, it is based on certain practices, e.g., focusing on the goal plan personification. As part of the goal setting and achievement assessment module in the reference model implemented in SuccessFactors (Figure 3 (a)), it is intended to increase the culture of communication between the manager and the employee, and to increase the employee's responsibility for achieving the goals through his/her involvement into the process of their setting. At forming goals by third parties, the process based on statistics preloading from the system and analyzing the goal performance for past periods, as implemented in the model shown on Figure 3 (b), a need to create extensions appears. A thorough pre-project analysis of customer's processes allows avoiding additional expenses associated with the unplanned writing of extensions.

It is noteworthy that the preliminary analysis is complicated by short terms for carrying out tenders for implementation, which provide for calculation of the total project cost based on limited project data, and high cost competition at the market of companies implementing CAIMS.

III. ANALYSIS OF OPTIMIZATION METHODS USED AT WORK AND EXAMPLES OF THEIR APPLICATION

This section deals with the task of planning the work of the CAIMS implementation team. Task solution may be carried out by various methods based on using: (1) Gant Chart (MS Project) – standard approach [7]; (2) imitation models [8, 3, 9]; (3) genetic algorithms [10-13]; (4) set of means including the imitation model (e.g., based on Petri networks) and task solution optimization with genetic algorithms [3]; (5) fuzzy sets [15].

The key requirements of the application solution being developed were:

- **Objectivity.** Development of criteria and formulas is carried out for each specific role (manager, hr-manager) in the project but not for an employee holding this position. Such objectification allows minimizing costs of model changes in the event of employees' dismissal or rotation. Moreover, this approach can be considered successful for companies building a project team or attracting external experts to solve the task.
- **Transparency.** Basing on models from recognized sources.
- **Intelligibility.** Detailed mathematical model description.
- **Dynamicity.** Ability to change the solution at a slight input data change (change in daily work time or the introduction of additional restrictions on incoming data).
- **Compliance with the Russian legislation** [16].

IV. SETTING THE TASK OF SEARCH OF OPTIMAL DISTRIBUTION OF PROJECT TEAMS' RESOURCES

The task considered in this article is related not to the production workshop operation but to the product introduction, and it is required to consider the features associated exactly with this type of activity when solving it. The articles [17, 18] describe the models and solve the tasks of resource allocation, including the situation when several criteria are present. The proposed model offers 3 minimization parameters: HR costs C , project time costs T_ϕ and possible risks R . The coefficients of these parameters importance are k_C , k_{T_ϕ} , k_R respectively:

$$Cost = k_C C + k_{T_\phi} T_\phi + k_R R \rightarrow \min \quad (1)$$

At this, if i is the task ordinal number, I - total number of tasks in the project, then variables C , T , R may be defined as follows:

$C = \sum_{j=1}^I c_j$ – total HR cost under each project task c_j incurred by the company implementing CAIMS;

j – task sequence number;

J – total number of tasks in the project;

$T_\phi = \sum_{j=1}^J T_j$ – the time of execution (implementation) of the

project, in the case of sequential execution of tasks, is defined as the sum of all the time spent on solving individual tasks within the project.

R – total risks, some of which are associated with additional resource costs and lengthening the time to the project, are described as follows:

$$R = k_{Co} R_{Co} + k_{Eo} R_{Eo} + k_{To} R_{To} \quad (2)$$

R_{Co} - risks related to internal resources;

k_{Co} - their importance coefficient;

R_{To} - risks related to the project terms;

k_{To} - their importance coefficient;

R_{Eo} - risks related to mobilization of external resources;

k_{Eo} - their importance coefficient.

Factors able to affect the model are given in Table 2.

TABLE II. RISK FORMING FACTORS

Risk type	Factors
R_{Co}	- Appearance of standard project tasks not planned at the project initial stages (e.g., the need of conducting additional trainings for the users) - Accounting of a disease and the need of developing internal resources causing the paid delays in work
R_{To}	- Technical problems causing temporary losses: service provider's technical problems, communication problems - Problems of uneven loading of employees, lack of real tasks
R_{Eo}	- Appearance of non-standard project tasks not planned at the project statement stage, for development of additional elements, optimization of separate processes

As practice of project implementation shows (Table 1), the actual period of project implementation often exceeds its planned estimate. This is especially clearly expressed for companies that do not pre-optimize the processes but automate their processes "as is". In such projects, a higher number of additional elements are developed, which, in its turn, leads to the need to mobilize additional resources, including external ones, and increases the project terms considerably.

On the basis of statistics, the factor of disease and the need to develop domestic resources can be taken into account. According to Rosstat (Federal Service of State Statistics) data for 2017, Russian people were ill on average for no more than 8 days per month, while some of them continued to attend work (up to 80%). For a young collective (20-35 years old), the average number of full medical leaves per month can be estimated as 1 out of 20 working days [19].

For a modern IT company, it is considered a good courtesy to provide its employees with at least 2 weeks a year for training and professional development. At this, not all the expenses for training consultants pay off: according to Antal Russia [20], the labour turnover in IT and telecom spheres in Russia

at 2018 is 11%. Therefore, the company implementing CAIMS loses about $0.11(Z \sum_{i=1}^l p_i)$ every month, where Z is the cost of training one employee per month, $p_1, p_2, p_3, \dots, p_l$ - number of team members for each type of staff on staff.

An additional model limitation is the need to complete the project on time:

$$T_\phi - T_{II} \rightarrow 0 \quad (3)$$

where T_{II} - planned project completion time.

In articles [8, 21], sequential operation was considered, and restrictions were imposed that one operation will start no earlier than the previous one ends. In the model offered, parallelization of task execution and their joint execution are considered. The model provides for the accounting of the following features (restrictions):

A. Accounting of the task implementation sequence.

Accounting of this feature in a mathematical model can be implemented through graph building. The example shown on Figure 4 illustrates the processes shown on Figure 2. The sequence of tasks correspond to the example of project number 5, table 1. Here, the graph source is the task N1 - the project start and obtaining the accesses. The graph sink point is the task N29 - the project completion (data deletion in the learning environment, according to figure 2). Other tasks are numbered, and sequence restrictions are defined for them.

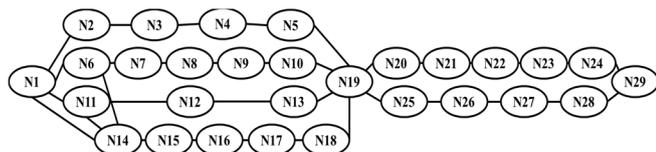


Fig. 4. Fragment of the graph describing the typical project tasks

B. Accounting of the possibility of simultaneous performance of one task by various consultants.

E.g., in [22], a modern project management method, Extreme Programming, is described, characterized by the product development and testing process carried out by a small team (from two to ten programmers) with daily goals setting. The authors of [22] assert that at such work organization, reducing task completion terms is possible. The simultaneous work of a small team on one unit is also typical for the agile (a family of "flexible" implementation methods). For the tasks and projects described above, pair programming can be used, suggesting the consultant's individual efficiency increase at paired activity].

On the other hand, the authors of [23] refer to the fact that when processing with IT products, specialized employees directly program for 55% of their working time. The remaining time is spent on communication with the management, colleagues, testers, designers and the customer. Moreover, when more than two consultants process one task, the need to perform these tasks in specialized applications (task management) appears, and these actions take time to be performed. Based on these provisions, Figure 5 shows the dependence of consultants' individual efficiency on the number of team members: from 100% efficiency at individual work to 55% efficiency at work of a team composed of 11 specialists. Piecewise linear approximation is applied.

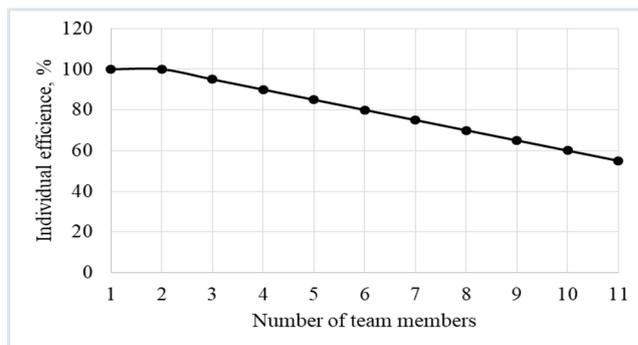


Fig. 5. Chart of the consultant's individual efficiency loss at teamwork K_n n - number of team members

C. Accounting of the possibility of simultaneous performance of various tasks by one consultant.

As shown in clause 2, the performance of one task by several consultants does not always lead to their work efficiency increase, e.g., performing several tasks by one consultant can reduce the excess communication cost. Therefore, such labor organization is permissible within the model.

D. Accounting of qualification / cost of various consultants.

According to [24], team management is a creative process and its efficiency depends on a large number of factors. The efficiency is considerably influenced by the culture of interaction in a team (along with its elements' individual abilities), and maintenance of informal relations between its members. As mentioned above, when working with the use of pair programming technique (for consultants of one qualification), the individual efficiency increase can be observed. On the other hand, the practice of the discussed product introduction shows that at work with younger colleagues, the more experienced ones spend time on mentoring (10-20%), explain the task specifics for longer time and due to this, their individual efficiency in solving the current task is considerably decreased, and that of their younger colleagues is acquired.

Table 3 shows the coefficient of the consultant's individual efficiency change at paired work: consultants and their efficiency are located horizontally, and their 'partner's' name is located vertically. The "X" sign is used to denote "forbidden" pairs: it is not recommended to involve two managers or architects within one project. The Table does not provide information about another type of employees, a project manager, whose load, as a rule, is estimated as 0.25 of the entire load throughout the work day in during all project.

TABLE III. INDIVIDUAL EFFICIENCY COEFFICIENT AT PAIRED ACTIVITY K_{i,j_2}

Resource type	Resource type					
	Business consultant	Architect	Senior consultant	Consultant	Junior consultant	Developer
Business consultant	X	1	0,8	0,8	0,8	0,8
Architect	1	X	0,8	0,8	0,8	0,8
Senior consultant	1,2	1,2	1,1	0,8	0,8	0,8
Consultant	1,2	1,2	1,2	1,1	0,8	0,8
Junior consultant	1,2	1,2	1,2	1,2	1,1	0,8
Developer	1,2	1,2	1,2	1,2	1,2	1,1

Features 2-4 may be accounted within the classical routing problem (Table 3). The project tasks are located horizontally, names of resources are located vertically, and cells show costs c_j for the j^{th} task that can be calculated by the formula:

$$c_j = \frac{1}{K_j} \sum_{i=1}^I m_i Pr_{ij} \quad (4)$$

where m_i is the i^{th} type resource hourly wage;

K_j – overall resource performance when working on the j^{th} task;

Pr_{ij} – resource performance of type i^{th} when working on the j^{th} task.

The hourly rate of resources is given in the second Table 4 column. The values of the planned performance and cost of resources correspond to the example of the project number 5, table 1. The resource performance is calculated as a task/day at its individual implementation. The permissible values of the number of employees reflect data for a number of projects of varying complexity and are taken into account as limitations.

$K_j(p_{1j}, p_{2j}, p_{3j}, \dots, p_{ij})$ is the personal efficiency function which depends on members of a team processing the task, with regard of the change in the consultant's individual performance K_{ij} at teamwork. The function accounts for the dependency given on Figure 5 (for teams of up to 3 persons) and data of Table 3 (for teams of 2 persons), dependence is described by formula 5:

$$K_j(p_{1j}, p_{2j}, p_{3j}, \dots, p_{ij}) = \begin{cases} 1, n = 1; \\ \frac{\sum_{i=1}^2 m_i p_{ij} Pr_{ij}}{m_i p_{i_1} Pr_{i_1} / K_{i_1} + m_i p_{i_2} Pr_{i_2} / K_{i_2}}, n = 2; \\ K_{nj}(p_{1j}, p_{2j}, p_{3j}, \dots, p_{ij}), n \geq 3. \end{cases} \quad (5)$$

$p_{1j}, p_{2j}, p_{3j}, \dots, p_{ij}$ - the number of team members of each type when working on the j^{th} task.

$$n = \sum_{i=1}^I p_{ij}^0, p_{ij} \neq 0 - \text{the number of } p_{ij} \text{ not equal } 0.$$

The sign 'X' denotes the forbidden positions.

TABLE IV. REPRESENTATION OF MODEL CRITERIA IN THE FORMAT OF THE CLASSIC ROUTING PROBLEM

Resource type	Options						Admissible number
	Daily wage, ye	Setting	Solution agreement	Document writing	Iterations	...	
Business consultant	4	X	$4K_i / 20$	X	X		0-1
Architect	3,5	$3,5K_i / 10$	$3,5K_i / 20$	$3,5K_i / 5$	$3,5K_i / 10$		1
Senior consultant	3	$3K_i / 15$	$3K_i / 30$	$3K_i / 7,5$	X		0-1
Consultant	2	$2K_i / 20$	$2K_i / 40$	$2K_i / 10$	X		0-2

Junior consultant	1	$K_i / 25$	X	$K_i / 12,5$	X		0-2
Developer	2,5	X	X	X	$2,5K_i / 25$		0-2

Accounting of feature 1 (unit integration sequence) jointly with the calculation performed in the course of solving the "routing problem" described above for labor resource allocation will allow assessing the project duration with regard of the task solving sequencing.

V. SELECTION OF A METHOD OF SOLVING THE TASK OF OPTIMAL PROJECT TEAMS' RESOURCES ALLOCATION

The genetic algorithm method was selected for the given task solution. To solve this problem, the genetic algorithm method was chosen, since this method was widely used [10–14] for solving problems of project management optimization. At the same time, other works did not simultaneously take into account all the conditions and possibilities considered in the model described in this article.

Basic genetic algorithm terms and their physical meaning in the task implemented are given in Table 5 [25, 26].

TABLE V. BASIC GENETIC ALGORITHM TERMS AND THEIR USE IN THE TASK SOLUTION.

Term	Definition	Use in the task
Gene	Chromosome atom element	With the purpose of involvement of each resource type into the solution of the j th task, $j = 1, \dots, J$. is expressed as $x_j = \{p_1, p_2, \dots, p_I\}$, where p_1, p_2, \dots, p_I - the contribution of time of each of I types of employees
Chromosome	Gene array	Task solution set $X = \{x_j\}$, where x_j is a gene corresponding to the j th task.
Fitness function	Chromosome adaptation coefficient	Superposition of the project costs which are determined on the basis of the cost of work of employees of different types, their productivity and the degree of involvement into the task and time limitations. Task solution reduces to minimization of the fitness function $Cost \rightarrow \min$.
Genetic operators	Crossing and mutation operators. Crossing-over is the crossing of two species. Mutation is the intended artificial change of the specific genes in the species chromosome	The arithmetical crossing-over was used in the task: if $Cost(K_{parent1}) \leq Cost(K_{parent2})$, then: $K_{child} = \alpha k_{parent,1} + (1 - \alpha)k_{parent,2}$; $\alpha = 0,9$, $K_{parent1}, K_{parent2}$ - parent chromosomes, and K_{child} - daughter chromosome. The rearrangement of several random gene positions was performed as the mutation. The following operators were used as the mutation: (1) rearranging of several chromosome positions (determining the share of resources occupied by separate tasks), (2) assigning null values to several cells (this mutation increased the task convergence rate considerably)

It is convenient to use the share of involvement of each resource type into each task as genes. At this, the time of the j^{th} task execution can be calculated univalently on the basis of individual efficiencies of the employees participating in the task (Table 3 and Figure 5) and their individual productivities (Table 4), depending on the number of members:

$$\Delta\tau_j = \sum_{i=1}^I \frac{1}{K_j Pr_{ij} p_{ij}} \quad (6)$$

Each task execution time is used in the calculation of restrictions on the sequence of their execution and in the final solution analysis.

At this, the percentage of employees' involvement is limited by the number of employees of each type in the project team: the project team may include several consultants and developers, and, as a rule, one architect and one business consultant.

At the model determination, it was considered that the employee was loaded daily for p_i of his/her performance during the task processing; at this, the load remains even during teamwork. Permissible workloads per employee per day were determined in increments of 0.25 and assumed values: 0, 0.25 - a quarter, 0.5 - half, 0.75 - three quarters, 1 - entirely. Loading multiple employees of the same type is indicated by values above 1 with the same step (1.25; 1.5; ...). The allowable values for the team members are given in Table 4.

Solution of the task project execution time minimization was carried out by calculating the critical path [27]. At this, the optimization task was reduced to minimizing the spare time for performing the tasks presented in the column on Figure 4, i.e., to minimizing the downtimes of the team involved into the project. The weights (lengths) of the graph edges are determined by the execution time of the j^{th} task. The graph edges were calculated on the basis of the formula (6).

A metric of optimal distribution of project team resources is the fitness function. It' is determined as follows:

$$Cost = k_C \sum_{j=1}^J \frac{1}{K_j (p_{1j}, p_{2j}, p_{3j}, \dots, p_{ij})} \sum_{i=1}^I m_i p_{ji} Pr_{ji} + k_{\Delta T} \Delta T \quad (7)$$

ΔT – the sum of the differences of the minimum and the maximum possible start of the execution of each graph node.

$k_{\Delta T}$ – ΔT parameter importance coefficient.

The flow-chart of task solution is given on Figure 6.

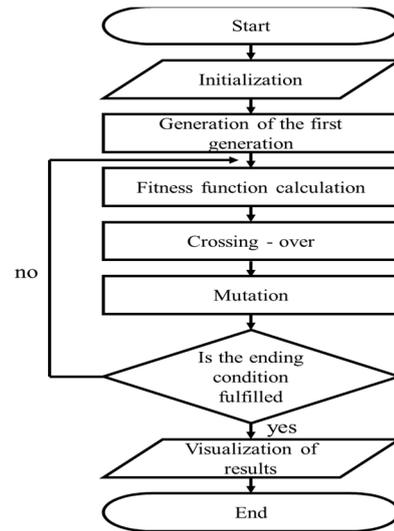


Fig. 6. Flow-chart of task solution by the genetic algorithm method Sequence sequence of operations: (1) Initial parameters are set (restrictions on employees of different types, coefficients of significance of time and cost variables for the project are determined); (2) The initial chromosome population is set; (3) The adaptation of the current population elements is assessed; (4) Chromosome crossing for new population creation is applied; (5) The mutation is applied; (6) The termination condition set in the form of a limit on the number of generations is verified; (7) The results are visualized with the use of the Gant Chart.

The solution of the problem was carried out in the Matlab software environment. Using the developed software, the results and graphs in the following sections are implemented. For the selected data volume, the solution was found in less than 100 iterations.

VI. RESULTS OF SOLVING THE TASK OF SEARCH OF OPTIMAL DISTRIBUTION OF PROJECT TEAMS' RESOURCES

The maximal number of employees of one type per project was set as the initial condition. Therefore, task solution assumed the selection of optimal executors of each task from total number of available employees, based on their qualification and cost. For the real process of the project team optimization, limiting the sample of employees involved into one project, i.e., the distribution of consultants of various types between various projects, is of interest.

Figure 7 shows the main main characteristics of the solution to the problem when changing the coefficients of significance of the parameters of time and team cost.

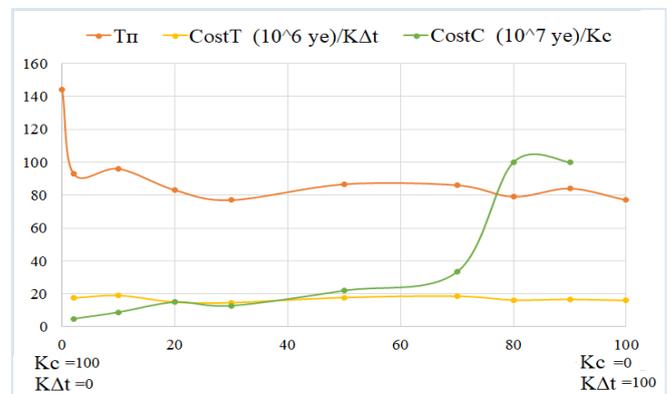


Fig. 7. The dependence of the main parameters of the coefficients of significance criteria: $CostC$ – the first term from formula (7); $CostT$ – the second term from formula (7); T_{π} – estimated time to complete the last task.

Thus, Figure 8 shows the distribution of resources throughout the project.

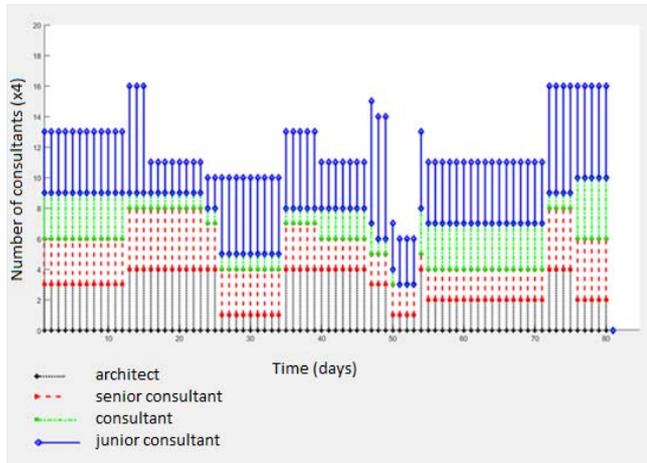


Fig. 8. Resource distribution on project

VII. PRACTICAL IMPORTANCE OF THE TASK OF OPTIMAL DISTRIBUTION OF PROJECT TEAMS' RESOURCES

In order to solve the set task, the use of the following parameters is suggested, beside the criteria of providing the set time for the task performance and the minimal project cost for the functioning of an enterprise the main kind of activity of which is the implementation of CAIMS and arranging the work of the project teams, the use of the following parameters:

- Maximal, minimal and average number of specialists of a given qualification, involved into the project implementation;
- Deviation of the current number of specialists of the given qualification involved into the project implementation, from the average value;
- Average squared displacement of the current number of specialists of the given qualification involved into the project implementation, from the average value.

Table 6 shows the value of the mathematical expectation of the consultants involved and the maximum possible parameters within the project under study.

TABLE VI. NUMBER OF CONSULTANTS ON ONE PROJECT

Resource type	Options				
	Minimum number	Maximum number	Possible number of consultants on the project	Mathematical expectation number of consultants	Average squared displacement of consultants
Business consultant	0	0	0	0	0
Architect	0	1	1	0.72	0.17
Senior consultant	0	1	1	0.73	0.27
Consultant	0	1	1	0.75	0.21
Junior consultant	0.25	1.75	2	1.13	0.31
Developer	0	0	0	0	0
Sum			5	3.33	

According to table 1 to example 5, the project under study was made by 5 consultants for 3.5 months. So we see the sum of the mathematical expectation number of consultants is 3.33. Thus it is possible to reduce the number of consultants in the project with more accurate planning.

When resources are not involved, we can use them in another project. For the project under study, the most sought-after employees are junior consultants, which may be due to the relatively low cost. The solution of the problem can be used to optimize the composition of the implementation team, optimize the cost of the project, determine the relationship between the duration of the project, its cost and the composition of the implementation team, detail the team work schedule. So, on fig. 9 as a solution to the problem, we see the classical Gantt chart, where the time to complete each task is divided into proportional parts between employees who participate in its implementation.

It is important that the proposed solution involves the work of consultants for up to 4 months, which corresponds to the initial requirement of the project under study.

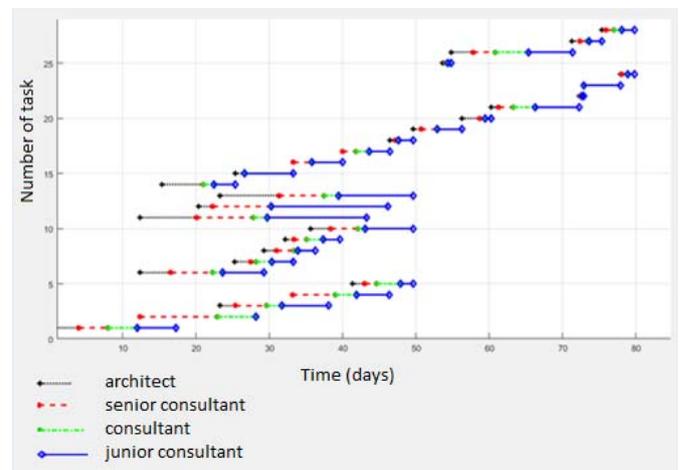


Fig. 9. An example of solving the problem

Labor intensity (months of consultants' work) according to the results ($3.33 \times 3.8 = 12.6$) is less than actual labor intensity ($5 \times 3.5 = 17.5$, table 1). Therefore, we can say that with the help of the developed product we can achieve team optimization.

VIII. CONCLUSIONS

We was developed a model of the project team optimization that takes into account a number of restrictions: accounting of the task implementation sequence, accounting of the possibility of simultaneous performance of one task by various consultants, accounting of the possibility of simultaneous performance of one task by various consultants, accounting of qualification / cost of various consultants.

The simultaneous inclusion of these conditions makes it possible to predict the work of the project team and calculate the costs for it in a future project more accurately than in existing models that take into account only part of these restrictions.

Created software allowed to solve the problem with the help of a genetic algorithm. The results indicate that it is possible to optimize the team and its costs using the developed tool.

The advantage of this solution is the absence of the need for additional infrastructure, the solution can be entered into the traditional planning system.

Currently, only two criteria were considered in solving the problem, a project for one module and a small team. The main feature of the solution is its ability to scale, and therefore work with the project pool, a significant team, and the added use risks as criteria. Solution of the task of optimizing the project team resources offered in this paper can be adapted to other tasks of automating the HR resource allocation management.

IX. REFERENCES:

- [1] V.A. Semenov, A.S. Anichkin, S.V. Morozov, O.A. Tarlapan, V.A. Zolotov "A complex method of scheduling for complex industrial programs taking into account the space-time constraints", Proceedings of the Institute Systems Programming RAS, ISSN: 2079-8156, 2014.
- [2] V.V. Muromtsev, A.V. Muromtseva "Human information space in conditions of state-of-the-art virtual communications", Components of Scientific and Technological Progress, 2014.
- [3] A.N. Sochnev "Production system resource allocation of the using Petri networks and the genetic algorithm", UBS. - 2012. - N 39.
- [4] N.P. Pilikov "AFM: Scheduler 1/11 Program for drawing up non-standard and optimal work schedules staff", AFM-Lab LLC, 2018.
- [5] A.V. Markevich "The current state of computerization of business processes", Collection of research articles on the results of the international research and practical conference, 2018, №6.
- [6] O.S. Reznikova "The application of "cloud" technologies in the management of the personnel of the organization", Science symbol, 2015.
- [7] O.M. Yanbulatova, S.A. Gordienko, A.E. Enis, T.G. Kirska "Software for project management", Mining information and analytical bulletin (scientific and technical journal), 2007.
- [8] N.D. Pronichev, V.G. Smelov, V.V. Kokareva, A.N. Malykhin "Simulation modeling of the machining workshop production system", News of the Samara Research Center of the Russian Academy of Sciences, 2013, №6-4, T.15.
- [9] A.V. Mishchenko, M.V. Mogilnitskaya "Dynamic model of production resources and circulating capital management in industrial logistics", INFRA-M, 2015.
- [10] R. Nedzelsky "Human Resources Allocation in Project Management", Conference: IBIMA, At Milan, Italy, Volume: 27. 2016.
- [11] M. Karova, G. Todorova, I. Penev, M. Todorova "Managing Project Activities System using Genetic Algorithm", Department of Computer Science and Engineering Technical University of Varna, 2015.
- [12] J. Park, D. Seo, G. Hong, D. Shin, J. Hwa, D. Bae "Practical Human Resource Allocation in Software Projects Using Genetic Algorithm", Department of Computer Science, 2012.
- [13] P.V. Afonin, O.V. Kokshagina, V.V. Naumenko "Hybrid Genetic Algorithms for the Task of Scheduling a Project", News of the Southern Federal University. Technical Sciences, 2008.
- [14] C. K. Chang, H. yi Jiang, Y. Di, D. Zhu, and Y. Ge. "Timeline based model for software project scheduling with genetic algorithms", Information and Software Technology. 50(11). – 2008.
- [15] D.A. Gradusov, Ie.S. Avdeeva, E.A. Ulanov "The use of fuzzy sets for assessing the economic efficiency of corporate information systems implementation projects", Economic analysis: theory and practice, 2012.
- [16] "Labor Code of the Russian Federation", Propaganda: OmegaL, 2002.
- [17] G.V. Lukin "Mathematical simulation of resource allocation tasks based on risk minimization", Scientific Library of theses and abstracts, 2005.
- [18] Ie.P. Rostova "Setting the task of dynamic programming for the distribution of risk management tools at the enterprise", News of the Samara Research Center of the Russian Academy of Sciences, 2013.
- [19] "Russia in figures", Federal Service state statistics, 2018.
- [20] M. Germershausen "Labor Market Survey and Salary Survey", Antal Russia, 2018.
- [21] A.V. Lagarnikova "The principle of data processing arrangement on the example of a five-speed conveyor", IV International Research and Practical Conference 'Students scientific community: interdisciplinary research., 2016.
- [22] J. Hank Rainwater "Herding Cats", Peter. - 2011. - p. 217-219.
- [23] W. Mickey "Mantle Managing the Unmanageable", 2012.
- [24] "Harvard Business Review of Top 10 articles for team management", Alpina, Moscow, 2017.
- [25] J.H. Holland "Adaptation in Natural and Artificial System", The MIT Press. - 1992.
- [26] M. Mitchell "An introduction to genetic algorithms", London: MIT Press. – 1999.
- [27] E.Yu. Shakhova "Search for critical paths in the graph", Proceedings of the Bratsk State University. Series: Natural and Engineering Sciences. - 2008. T.1.

Evaluating the length of distinguishing sequences for nondeterministic Input/Output automata

Igor Burdonov
Software Engineering
department

Ivannikov Institute for System
Programming of RAS
Moscow, Russia
igor@ispras.ru

Alexandr Kossachev
Software Engineering
department

Ivannikov Institute for System
Programming of RAS
Moscow, Russia
kos@ispras.ru

Nina Yevtushenko
Software engineering
department

Ivannikov Institute for System
Programming
Moscow, Russia
evtushenko@ispras.ru

Alexey Demakov
Software engineering
department

Ivannikov Institute for System
Programming
Moscow, Russia
demakov@ispras.ru

Abstract—Distinguishing sequences are used in model based mutation testing in order to distinguish the specification from its mutants that usually represent critical implementation faults. In this paper, we consider distinguishing sequences for Input/Output automata when a sequence of inputs can be applied before getting any response or a sequence of output responses from an implementation under test. We propose a technique for deriving an r -distinguishing trace, i.e. a distinguishing trace with respect to the trace inclusion (quasi-reduction) relation, and obtain the least and upper bounds on the length of a shortest r -distinguishing trace showing that the exponential upper bound with respect to the number of states of the specification automaton is reachable; the results are then adapted for a proper case of Input/Output automata when each input is followed by an output, i.e., for Finite State Machines.

Keywords—Input/Output automata, quasi-reduction relation, r -distinguishing trace

I. INTRODUCTION

Test generation with guaranteed fault coverage is an important issue in developing complex critical systems (see, for example, [1]) and guaranteed fault coverage immediately asks for involving formal models. Finite transition systems are widely used for deriving tests and there are a number of methods [2] for deriving test suites with guaranteed fault coverage for Finite State Machines (FSMs) when each input is followed by an output. However, the above FSM model is not always appropriate and sequences of inputs can be applied before getting any response or a sequence of output responses; this situation can be adequately handled by the use of so-called Input/Output(I/O) automata [3]. Nevertheless, all the methods developed for such automata usually return infinite tests when talking about the ‘black-box’ testing model [4]. In a number of cases when critical faults could be enumerated, a test suite can be derived as a set of distinguishing sequences for specification and mutant I/O automata. In this case, a technique for deriving an appropriate preset or an adaptive distinguishing sequence for two I/O automata has to be elaborated and the complexity of a corresponding test suite has to be evaluated. For FSMs there are many publications how to derive such sequences but we are not aware of these results for I/O automata.

In this paper, we consider a variation of the well known *ioco* relation [4] but as we consider automata that not necessary are input complete we modify *ioco* as a quasi-reduction relation (similar to FSMs [5]). Automaton A is not a quasi-reduction of automaton B if there exists a trace defined at both automata such that the set of outputs after this trace of automaton A is not a subset of that of automaton B and propose a technique how to check whether this relation holds. If the automaton B is deterministic then the obtained criterion describes necessary and sufficient conditions for checking the quasi-reduction relation. However, if the automaton B is nondeterministic then the conditions become only sufficient. Moreover, as we consider automata which not necessary are input complete and we do not observe states when testing, only traces for which an input is defined at any state after a corresponding prefix are considered as distinguishing test cases and such traces are called permissible. In order to completely check the quasi-reduction relation when B has a trace that takes the automaton from the initial state to two different states, the B has to be determinized; however, the deterministic equivalent of B is a bit different from the ordinary [6] as it contains only permissible traces and in this paper, we also evaluate the length of such trace obtaining lower and upper bounds for r -distinguishing traces depending on the number of states of both automata.

The rest of the paper is structured as follows. As usual, Section 2 contains preliminaries while a technique for deriving a distinguishing trace together with the least bound of such trace is presented in Section 3. Section 4 shows that the exponential upper bound on length of a shortest r -distinguishing trace with respect to the number of states of the specification automaton is reachable if the latter can be nondeterministic; the results are adapted for FSMs in Section 5. Section 6 concludes the paper.

II. PRELIMINARIES

An I/O automaton, simply an automaton throughout this paper, is a 5-tuple $\mathcal{S} = (V(\mathcal{S}), X, Y, E(\mathcal{S}), s_0)$ where $V(\mathcal{S})$ is a finite nonempty set of states with the initial state s_0 , X is a finite nonempty set of inputs, Y is a finite nonempty set of outputs, X

$\cap Y = \emptyset$, $E(\mathcal{S}) \subseteq V(\mathcal{S}) \times (X \cup Y) \times V(\mathcal{S})$ is a set of transitions. Sometimes, we refer to inputs and outputs as to *actions*. According to the above definition, we consider automata without the nonobservable action.

For $s, s' \in V(\mathcal{S})$ and $z \in (X \cup Y)$, we use the following notations:

$$s \xrightarrow{z} s' \stackrel{\text{def}}{=} (s, z, s') \in E(\mathcal{S}),$$

$$s \xrightarrow{z} \stackrel{\text{def}}{=} \exists s' \in V(\mathcal{S}) (s, z, s') \in E(\mathcal{S}).$$

If there are no transitions at a state under outputs then we add a loop labeled with ‘output’ δ [4] that is not in the set Y :

$$E_\delta(\mathcal{S}) = E(\mathcal{S}) \cup \{a \xrightarrow{\delta} a \mid a \in V(\mathcal{S}) \ \& \ \forall y \in Y \nexists b \ a \xrightarrow{y} b\}.$$

Such an augmented automaton \mathcal{S} is denoted as $\mathcal{S}_\delta = (V(\mathcal{S}), X, Y, E_\delta(\mathcal{S}), s_0)$, and a trace in \mathcal{S}_δ is a *S-trace*¹ of \mathcal{S} . If the contrary not explicitly stated then a trace denotes an *S-trace*.

Input $x \in X$ is a *defined input* at state $s \in V(\mathcal{S})$ if there is a transition $s \xrightarrow{x}$, i.e., $\exists s' \in V(\mathcal{S}) (s, x, s') \in E(\mathcal{S})$. Input $x \in X$ is *defined after a trace* if this input is defined at each state reached after this trace. An automaton is *input-complete* if every input is defined at every state.

Given a trace μ and state s , μ is a *permissible trace* at state s if each input in μ is defined after the prefix that directly precedes this input. Given a permissible trace μ at state s , as usual, s -*after*- μ is the set of states where μ can take the automaton from state s . If μ is a *permissible trace* at the initial state of \mathcal{S} then instead of s_0 -*after*- μ we sometimes write \mathcal{S} -*after*- μ . In this paper, we assume that two automata can be distinguished only by a trace that is permissible at the initial states of both automata; moreover, we also assume that if after an input no outputs are expected then δ is a corresponding output.

An automaton is *observable* if at each state at most one transition is defined for each action.² Given an observable automaton, the set of states s -*after*- μ is either empty or is a singleton. Given a nonobservable automaton, the set s -*after*- μ can have several states.

We also use the following notation: the set of outputs at a state s is the set of defined outputs at this state:

$$\text{outs}(s) \stackrel{\text{def}}{=} \{y \in Y \mid s \xrightarrow{y}\}. \text{ If no outputs are defined at state } s \text{ then } \text{outs}(s) \stackrel{\text{def}}{=} \{\delta\}.$$

$$\text{For a subset } B \subseteq V(\mathcal{S}) \text{ we have } \text{outs}(B) \stackrel{\text{def}}{=} \cup \{\text{outs}(s) \mid s \in B\}.$$

An automaton $\mathcal{A} = (V(\mathcal{A}), X, Y, E(\mathcal{A}), a_0)$ is a *quasi-reduction* of the automaton $\mathcal{S} = (V(\mathcal{S}), X, Y, E(\mathcal{S}), s_0)$ if for each trace σ that is permissible in both automata it holds that $\text{out}_{\mathcal{A}}(\mathcal{A}\text{-after-}\sigma) \subseteq \text{outs}(\mathcal{S}\text{-after-}\sigma)$.

If the automaton $\mathcal{A} = (V(\mathcal{A}), X, Y, E(\mathcal{A}), a_0)$ is not a quasi-reduction of the automaton $\mathcal{S} = (V(\mathcal{S}), X, Y, E(\mathcal{S}), s_0)$, then there exists a trace σ that is permissible in both automata such that $\text{out}_{\mathcal{A}}(\mathcal{A}\text{-after-}\sigma) \not\subseteq \text{outs}(\mathcal{S}\text{-after-}\sigma)$. This trace σ is called an *r-distinguishing trace*.

III. THE UPPER BOUND OF AN \mathcal{r} -DISTINGUISHING TRACE CASE DERIVATION

Let an automaton \mathcal{A} be not a quasi-reduction of \mathcal{S} , and σ is an *r-distinguishing trace*. For a trace σ , consider sequences of pairs $(a_j, \mathcal{S}\text{-after-}\sigma_j)$ where σ_j is a prefix of σ of length j , $j = 0, \dots, |\sigma|$, and $a_j \in (\mathcal{A}\text{-after-}\sigma_j)$, i.e., a_j is a state of \mathcal{A} reachable after trace σ_j . If σ is a shortest *r-distinguishing trace* with this property then there exists at least one sequence of pairs where all the pairs are pairwise different. Since the number of such pairs does not exceed the product of $n = |V(\mathcal{A})|$ and $2^k - 1$ where $k = |V(\mathcal{S})|$, the length of such sequence does not exceed $n(2^k - 1)$, and thus, the length of a shortest *r-distinguishing trace* is not bigger than $\mathbf{O}(n2^k)$.

Given an automaton \mathcal{A} and an observable automaton $\mathcal{S} = (V(\mathcal{S}), X, Y, E(\mathcal{S}), s_0)$ over the same alphabets, in order to derive a set of permissible traces of both automata, the product $\mathcal{A} \cap \mathcal{S}$ of automata can be constructed. States of the product are pairs of states of the automata, a transition is defined at a state if it is defined at both states.

Proposition 1. Given an automaton \mathcal{A} and an observable automaton \mathcal{S} over the same alphabets, \mathcal{A} is not a quasi-reduction of \mathcal{S} if and only if the product $\mathcal{A} \cap \mathcal{S}$ has a state (a, s) , $a \in V(\mathcal{A})$, $s \in V(\mathcal{S})$, such that the state is reachable from the initial state via a permissible trace at the initial states of both automata and some output is defined at state a while not being defined at state s .

Indeed, let (a, s) be a state with the above features reachable from the initial state via a trace μ . Since the automaton \mathcal{S} is observable, s is the only state of the automaton \mathcal{S} reachable by μ and the latter immediately implies $\text{out}_{\mathcal{A}}(\mathcal{A}\text{-after-}\sigma) \not\subseteq \text{outs}(\mathcal{S}\text{-after-}\sigma)$. On the other hand, if each permissible trace at the initial states of both automata takes the product $\mathcal{A} \cap \mathcal{S}$ to a state (a, s) such the set of outputs at state a is a subset of that at state s then by definition, \mathcal{A} is a quasi-reduction of \mathcal{S} .

It is well known how to construct the product of two automata over the same alphabet of actions and thus, Proposition 1 provides necessary and sufficient conditions for checking whether one automaton is a quasi-reduction of another observable automaton. If the product has a state (a, s) with the above features then a permissible trace μ that takes the product to this state is an *r-distinguishing trace*. Given an automaton \mathcal{A} with n states and an observable automaton \mathcal{S} with k states, the product $\mathcal{A} \cap \mathcal{S}$ has at most nk states, and thus, length of a shortest *r-distinguishing trace* is not bigger than $\mathbf{O}(nk)$.

¹Suspension trace

²Sometimes such an automaton is called deterministic [6]. However, we save this notion for an observable automaton where at each state at most one output is defined.

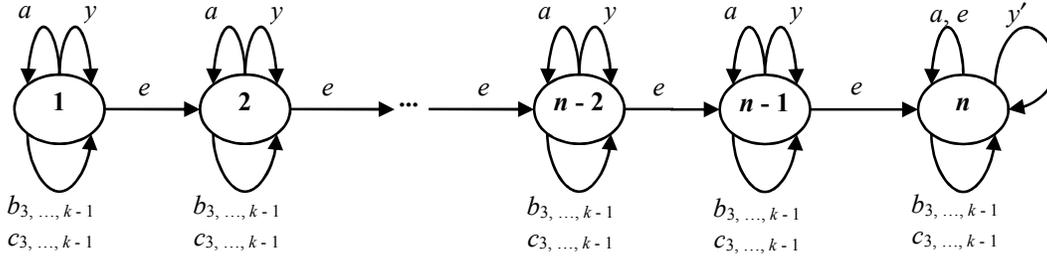


Fig. 1. Automaton A_n , $n \geq 2$

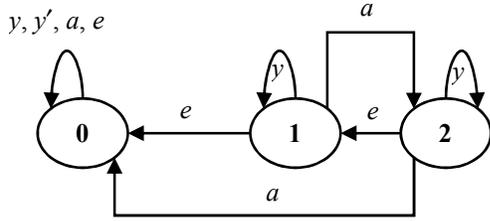


Fig. 2. Automaton S_3

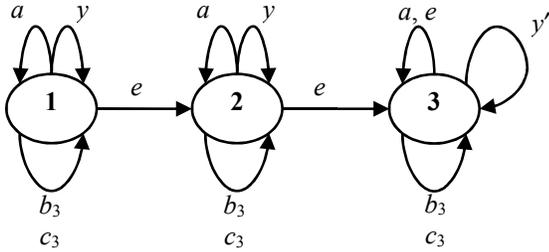


Fig. 3. Automaton A_3

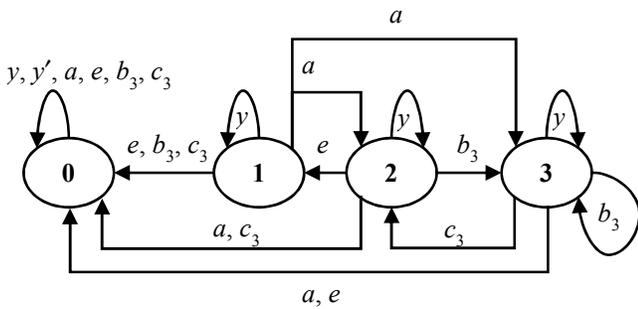


Fig. 4. Automaton S_4

If automaton $\mathcal{S} = (V(\mathcal{S}), X, Y, E(\mathcal{S}), s_0)$ is not observable, we define a power-automaton $\mathcal{A}(\mathcal{S})$ over the same alphabets X and Y . We use non-empty subsets of the set $V(\mathcal{S})$ as states of $\mathcal{A}(\mathcal{S})$, i.e., $V(\mathcal{A}(\mathcal{S})) = 2^{V(\mathcal{S})} \setminus \{\emptyset\}$, and the initial state of $\mathcal{A}(\mathcal{S})$ is $\{s_0\}$. In the automaton $\mathcal{A}(\mathcal{S})$, there is a transition $A \rightarrow x \rightarrow B$ under input $x \in X$ if and only if x is a defined input at each state $a \in A$ of \mathcal{S} , and $B = \{b \mid \exists a \in A \ a \rightarrow x \rightarrow b\}$. In $\mathcal{A}(\mathcal{S})$, a transition $A \rightarrow y \rightarrow C$

is defined under output $y \in Y$ if and only if in \mathcal{S} , a transition under this output is defined at least at one state $a \in A$, and $C = \{c \mid \exists a \in A \ a \rightarrow y \rightarrow c\}$.

Proposition 1'. Given automata A and $\mathcal{S} = (V(\mathcal{S}), X, Y, E(\mathcal{S}), s_0)$ over the same alphabets, A is not a quasi-reduction of \mathcal{S} if and only if the product $A \cap \mathcal{A}(\mathcal{S})$ has a state (a, δ) , $a \in V(A)$, $\delta \in V(\mathcal{A}(\mathcal{S}))$, such that the state is reachable from the initial state via a permissible trace at the initial states of both automata and some output is defined at state a while not being defined at the power-state δ .

Propositions 1 and 1' show the way how an r -distinguishing trace can be constructed if automaton A is not quasi-reduction of \mathcal{S} .

In the next section, we show that for every $k \geq 2$ and $n \geq 1$, there exist an automaton S_k with k states, $(2k - 2)$ inputs and two outputs and an input-complete automaton A_n with n states over the same input and output alphabets that is not a quasi-reduction of S_k such that a shortest r -distinguishing trace has the length $(n - 1)2^{k-2} = \Omega(n2^k)$.

IV. THE LOWER BOUND OF AN r -DISTINGUISHING TRACE

Theorem 2. For every $k \geq 3$ and $n \geq 1$, there exist an input-complete automaton S_k with k states, $(2k - 2)$ inputs and two outputs and an input-complete automaton A_n with n states over the same input and output alphabets that is not a reduction of S_k , such that a shortest r -distinguishing trace has the length $(n - 1)2^{k-2} = \Omega(n2^k)$.

Sketch of the proof. In order to prove the statement, we derive automata S_k and A_n for which the bound is reachable. The input alphabet $X = \{a, e, b_3, \dots, b_{k-1}, c_3, \dots, c_{k-1}\}$ has $2(k - 2)$ inputs while the output alphabet has two outputs, $Y = \{y, y'\}$. The set $V(S_k)$ of states of S_k is the set $\{0, 1, 2, \dots, k - 1\}$ and state 1 is the initial state.

Automaton S_k has the following transitions:

State 0: There are transitions under all inputs and all outputs to state 0;

State 1: There is a transition under input a to each state of the set $\{2, \dots, k - 1\}$; for each input $e, b_3, \dots, b_{k-1}, c_3, \dots, c_{k-1}$, there is a transition to state 0 while there is a transition to state 1 under output y ;

State 2: There is a transition under input e to state 1; for each input $b_j, j = 3, \dots, k - 1$, there is a transition from state 2 to state j under input b_j ; for each action $a, c_j, j = 3, \dots, k - 1$, there is a

transition from state 2 to state 0 while there is a transition to state 2 under output y ;

State $j, j = 3, \dots, k-1$: there are transitions to state j under inputs $b_3, \dots, b_j, c_3, \dots, c_{j-1}$ and transitions to states $2, \dots, (j-1)$ under input c_j ; there are transitions to state 0 under inputs $b_{j+1}, \dots, b_{k-1}, c_{j+1}, \dots, c_{k-1}$ to state 0; there is a transition to state 0 under inputs a and e and a transition to state j under output y .

Automaton A_n is shown in Fig. 1. Automata S_3, A_3, S_4 are shown in Figs. 2, 3, 4.

By direct inspection, one can assure that the automaton A_1 is distinguishable from $S_k, k \geq 3$, with the empty trace of length 0 and a shortest r -distinguishing trace for automata A_3 and S_3 is a trace $aeae$ of length $4 = (3-1)2^{3-2}$.

We first establish several statements about properties of automata A_n and S_k .

Proposition 3. A trace σ is an r -distinguishing trace of A_n with respect to S_k if and only if this trace takes an power-automaton $\mathcal{P}(S_k)$ from the initial state to any power-state without state 0 while taking the automaton A_n to state n .

Indeed, according to automata definitions, output y' can be produced at any power-state with state 0 and only at such power-state.

Proposition 4. Given a state j of $A_n, j < n$, a trace $\gamma \in \{a, b_3, \dots, b_{k-1}, c_3, \dots, c_{k-1}\}^* e$ takes the automaton A_n from state j to state $j+1, j = 1, \dots, n-1$, while taking the automaton from state n to state n .

The proof is a corollary to the fact that by definition, given a state j of $A_n, j \leq n$, a trace $\gamma \in \{a, b_3, \dots, b_{k-1}, c_3, \dots, c_{k-1}\}^*$ leaves the automaton at state j .

Due to the definition of the automaton S_k , the following statement holds.

Proposition 5. Given automaton S_k and a trace $\gamma \in \{b_3, \dots, b_{k-2}, c_3, \dots, c_{k-2}\}^*$ that takes the power-automaton $\mathcal{P}(S_k)$ from a power-state $\{2, \dots, k-2\}, k \geq 4$, to the power-state $\{2\}$ traversing power-states $D_1, \dots, D_{|\gamma|}$ without state 0, the trace γ takes the power-automaton $\mathcal{P}(S_k)$ from the power-state $\{2, \dots, k-2, k-1\}$ through the power-states $D_1 \cup \{k-1\}, \dots, D_{|\gamma|} \cup \{k-1\}$ each of which has state $(k-1)$.

By definition, a trace γ is empty in Proposition 5 when $k = 4$.

We first consider the case of $n \geq 2$ and $k = 3$. In this case, the automaton S_3 has only inputs a and e and by direct inspection, one can assure that a trace ae takes the automaton S_3 from state 1 to state 1 while taking the automaton A_n from every state $j \neq n$ to state $j+1$, i.e., the trace $(ae)^{n-1}$ takes the automaton S_3 from state 1 to state 1 while taking the automaton A_n from 1 to state n , and thus, is a r -distinguishing trace for A_n and S_3 . An input e after trace of the set $(ae)^*$ takes the automaton S_3 from state 1 to state 0 and input a after any trace of the set $(ae)^* e$ takes the automaton S_3 from state 2 to state 0; therefore, $(ae)^{n-1}$ is a shortest r -distinguishing trace for A_n and S_3 . This trace has length $(n-1)2^{k-2}$.

Let $n \geq 2$ and $k \geq 4$. We now use the induction on k in order to show that there is a trace of length 2^{k-2} that takes S_k from state 1 to state 1.

Induction base. If $k = 4$ then the trace a, b_3, c_3, e possesses the feature while traversing the following power-states: $\{1\} - a$

$- \{2, 3\} - b_3 - \{3\} - c_3 - \{2\} - e - \{1\}$. At any power-state of the trace, any other input takes the automaton to state 0 or to a power-state already traversed by the trace.

Induction assumption. Let for some $k < m$ hold that a trace $a\gamma, \gamma \in \{b_3, \dots, b_{k-2}, c_3, \dots, c_{k-2}\}$, takes the power-automaton $\mathcal{P}(S_k)$ from power-state $\{1\}$ to $\{2, \dots, k-2\}$ and from power-state $\{2, \dots, k-2\}$ to $\{2\}$ traversing power-states $D_1, \dots, D_{|\gamma|}$ without state 0 and length of this trace is $2^{k-2} - 1$, i.e., the trace $a\gamma e$ takes the power-automaton $\mathcal{P}(S_k)$ from power-state $\{1\}$ to $\{1\}$. We append the trace γ with $b_{k-1}c_{k-1}$, i.e., the trace $a\gamma b_{k-1}c_{k-1}$ of the power-automaton $\mathcal{P}(S_k)$ traverses power-states $\{2, \dots, k-1\}, D_1 \cup \{k-1\}, \dots, D_{|\gamma|} \cup \{k-1\}, \{k-1\}, \{2, \dots, k-2\}$ from state $\{1\}$ (Proposition 5). Correspondingly, the trace $a\gamma b_{k-1}c_{k-1}\gamma e$ takes the power-automaton $\mathcal{P}(S_k)$ from power-state $\{1\}$ to $\{1\}$ while taking automaton A_n to state 2 (Proposition 4).

Therefore, the trace $(a\gamma b_{k-1}c_{k-1}\gamma e)^{n-1}$ takes the power-automaton $\mathcal{P}(S_k)$ from power-state $\{1\}$ to $\{1\}$ while taking automaton A_n to state n , and due to Proposition 3, this proves the theorem statement.

V. EVALUATING LENGTH OF AN r -DISTINGUISHING TRACE FOR FINITE STATE MACHINES

The notion of a Finite State Machine (FSM) is very close to the notion of an I/O automaton. In fact, an FSM correspond to an I/O automaton where only inputs or outputs are defined at each state and each input is followed exactly by a sequence of outputs of length 1. Therefore, there are no races between inputs and outputs in FSMs and this fact makes this model very attractive for deriving test suites.

Formally, an initialized FSM is a 5-tuple $S = (S, X, Y, h_S, s_0)$ [7] where S is a finite non-empty set of states with the designated initial state s_0 , X and Y are input and output alphabets, and $h_S \subseteq S \times X \times Y \times S$ is the transition (behavior) relation. A transition (s, x, y, s') describes the situation when an input x is applied to S at the current state s . In this case, the FSM moves to state s' and produces the output (response) y . FSM S is nondeterministic [8] if for some pair $(s, x) \in S \times X$, there can exist several pairs $(y, s') \in Y \times S$ such that $(s, x, y, s') \in h_S$; otherwise, the FSM is deterministic. FSM S is observable if for every two transitions $(s, x, y, s_1), (s, x, y, s') \in h_S$ it holds that $s_1 = s_2$; otherwise, the FSM is nonobservable.

FSM S is *complete* if for each pair $(s, x) \in S \times X$ there exists $(y, s') \in Y \times S$ such that $(s, x, y, s') \in h_S$; otherwise, the FSM is *partial*. Given state $s \in S$ and an input $x \in X$, an input x is a *defined* input at state s if there exists $(y, s') \in Y \times S$ such that $(s, x, y, s') \in h_S$. Given an input sequence $\alpha = x_1 x_2 \dots x_k \in X^*$, α is a *defined input sequence* at state s if x_1 is a defined input at state s and for each $j = 2, \dots, k, x_j$ is a defined input at any state where input sequence $x_1 x_2 \dots x_{j-1}$ can take FSM S from state s .

In usual way, the behavior relation is extended to input and output sequences. Given states $s, s' \in S$, a defined input sequence $\alpha = x_1 x_2 \dots x_k \in X^*$ at state s and an output sequence $\beta = y_1 y_2 \dots y_k \in Y^*$, there is a transition $(s, \alpha, \beta, s') \in h_S$ if α is a defined input sequence at state s and there exist states

$s_1 = s, s_2, \dots, s_k, s_{k+1} = s'$ such that $(s_{j-1}, x_j, y_j, s_j) \in h_S, j = 1, \dots, k$. In this case, the input sequence α can take (or simply takes) the FSM S from state s to state s' . The set $outs(s, \alpha)$ denotes the set of all output sequences (responses) that the FSM S can produce at state s in response to a defined input sequence α , i.e. $outs(s, \alpha) = \{\beta: \exists s' \in S [(s, \alpha, \beta, s') \in h_S]\}$. The pair $\alpha \circ \beta, \beta \in outs(s, \alpha)$, is an *Input/Output (I/O) sequence* at state s ; if s is the initial state s_0 then the pair α/β is an *Input/Output (I/O) sequence* (or a *trace*) of the FSM S . Given states s and s' , the I/O sequence α/β can take (or simply takes) the FSM S from state s to state s' if $(s, \alpha, \beta, s') \in h_S$. Given FSMs $S = (S, X, Y, h_S, s_0)$ and $P = (P, X, Y, h_P, p_0)$, the *intersection* (or the *product*) $S \cap P$ is the largest connected submachine of FSM $= (S \times P, X, Y, f, s_0 p_0)$ where $(sp, x, y, s'p') \in f \Leftrightarrow (s, x, y, p') \in h_S \& (p, x, y, p') \in h_P$. The set $successor(s, \alpha \circ \beta)$ denotes the set of all states reachable from state s after applying the defined input sequence α when getting the output response β , i.e., given a defined input sequence α at state s , $successor(s, \alpha \circ \beta) = \{s': (s, \alpha, \beta, s') \in h_S\}$.

Given FSMs S and P , FSM P is a *quasi-reduction* of S if for each input sequence α defined at the initial states of FSMs S and P , it holds that $out_P(p_0, \alpha) \subseteq out_S(s_0, \alpha)$; otherwise, if there exists input sequence α defined at the initial states of FSMs S and P such that $out_P(p_0, \alpha) \not\subseteq out_S(s_0, \alpha)$, then P is not a quasi-reduction of S and α is a *r-distinguishing* (input) sequence. If both machines S and P are complete then the quasi-reduction relation reduces to the reduction relation: FSM P is a *reduction* of S if and only if for each input sequence α , it holds that $out_P(p_0, \alpha) \subseteq out_S(s_0, \alpha)$. In [8], it is shown that for two complete observable FSMs P with $n \geq 1$ states and S with $k \geq 1$ states, length of a shortest *r-distinguishing* sequence does not exceed nk and this bound is reachable for machines with a single input when n and k are relatively prime integers. If FSM S is not observable then an *r-distinguishing* sequence has length at most $n2^k$ but the reachability for this upper bound was not proven. Converting machines A_n and S_k from Section 4 into FSMs by replacing at each state every input by the pair input/output for the output defined at the state, the following statement can be established.

Theorem 6. For every $k \geq 3$ and $n \geq 1$, there exist complete FSMs S_k with k states, $2(k-2)$ inputs and two outputs, and a complete deterministic FSM A_n with n states over the same

input and output alphabets that is not a reduction of S_k , such that a shortest *r-distinguishing* sequence has the length $(n-1)2^{k-2} = \Omega(n2^k)$.

VI. CONCLUSION

In this paper, we are concerned about the complexity of test suites with guaranteed fault coverage when critical faults are enumerated and a test suite is derived as a set of distinguishing sequences of the specification and mutant I/O automata when a sequence of inputs can be applied before getting a response or a sequence of output responses from an implementation under test. We propose a technique for deriving an *r-distinguishing* trace of the specification and a mutant I/O automata, i.e., a distinguishing trace with respect to the trace inclusion (quasi-reduction) relation, and obtain the least and upper bounds on the length of a shortest *r-distinguishing* trace showing that the exponential upper bound with respect to the number of states of the specification automaton is reachable. The results are then adapted for a proper case of I/O automata when each input is followed by an output, i.e., for Finite State Machines. As further directions of our work, we are going to study other distinguishability relations especially those when adaptive input sequences can be used.

ACKNOWLEDGMENT

This work is partly supported by RFBR project N 19-07-00327/19.

REFERENCES

- [1] Mathur, A.: Foundations of Software Testing. Addison Wesley (2008)
- [2] Dorofeeva, R., El-Fakih, K., Maag, S., Cavalli, A., and Yevtushenko, N.: FSM-based conformance testing methods: A survey annotated with experimental evaluation. Inf. Software Technol., 52: 1286-1297 (2010)
- [3] N. Lynch and M. Tuttle. An introduction to Input/Output automata. CWI-Quarterly, 2(3): 219-246 (1989)
- [4] J. Tretmans.: A formal approach to conformance testing. The Intern. Workshop on Protocol Test Systems, 257-276 (1993)
- [5] Petrenko, A., Yevtushenko, N., Lebedev, A., Das, A. Nondeterministic State Machines in Protocol Conformance Testing. The Intern. Workshop on Protocol Test Systems: 363-378 (1993)
- [6] Hopcroft J.E., Motwani, R., and Ullman J.D.: Introduction to Automata Theory, Languages, and Computation. Addison-Wesley, second edition (2001)
- [7] Kam, T., Villa, T., Brayton, K. R., Sangiovanni-Vincentelli, A.: Synthesis of FSMs: Functional Optimization. Springer (1997)
- [8] Yevtushenko N., Petrenko A., Vetrova M.: Nondeterministic Finite State Machines: analysis and synthesis. Part 1: Relations and operations (in Russian). Publishers TSU, Tomsk (2006)

Voltage Regulation Analysis in Energy Transmission Systems Using STATCOM

1st Hamza Feza Carlak
Department of Engineering
Electrical and Electronics, University
of Akdeniz University,
Antalya, Turkey,
e-mail: fezacarlak@akdeniz.edu.tr
https://orcid.org/0000-0002-8561-4591

2nd Ergin Kayar
Department of Engineering
Electrical and Electronics, University
of Akdeniz University,
Antalya, Turkey,
e-mail: erginkayar07@gmail.com
https://orcid.org/0000-0002-7356-2165

Abstract— Due to the difficulties to insert new power transmission lines, dramatic increases in electricity demand, and proceeding the stability of power systems turns out to be the most critical and challenging problem. Modeling of the energy transmission system is performed for the pilot region to optimize controllable parameters of the power system to maintain energy quality at the most appropriate value. The simulation study is implemented for the Denizli western region energy transmission system with realistic values. The load flow analysis of the system has been carried out by using DigSilent power system analysis software. The control of voltage fluctuations of the national electric energy network system by the help of FACTS (Flexible Alternating Current Transmission Systems) technology is carried out with the proposed method using real national data for the pilot region, and the results are compared with the existing network system. The usage of STATCOM (Static Synchronous Compensators) controllers which are a part of a new control system called FACTS, and capacitor banks are compared for the dynamic power system model, and the results of each device are presented in the study. The STATCOM has a very high reaction time and operates in a wide range depending on their capacity and provides more flexible and safe operation. Moreover, the maximum load limits can be increased, and the control of the power system may be facilitated utilizing the STATCOM devices.

Keywords— Power Quality, STATCOM, Voltage Regulation, Reactive Power Control, Voltage Stability

I. INTRODUCTION

Flexible Alternating Current Transmission Systems are used to control the system thanks to the use of high-power electrical systems. In this context, determination of the location and size of the devices that provide essential benefits in power flow control, dynamic stability, continuous and transient state stability, transmission transfer capacity increase, voltage stability, reactive power control are of great importance both technically and economically. In the literature, although the primitive studies for FACTS (Flexible Alternating Current Transmission Systems) technology have been made, and the controllers have been classified and examined, the feasibility study for a real power system model that has been constructed with realistic values considering all the parameters for a real energy transmission system has not been performed yet. Dynamic operations in power transmission networks are essential for proper system planning and maintenance, and the results obtained to ensure that the proper steps are taken to prevent the system from

operating under unstable conditions that may eventually cause partial or complete collapse. The aim of this article is to evaluate the voltage stability of the pilot power system modelled as ten bus. The pilot region with realistic power and transmission line parameters belongs to the western region of Denizli Power System is located in the Western Mediterranean Region of Turkey. A method that determines the optimum placement and values of Flexible Alternating Current Transmission Systems devices in terms of bus voltage changes and line capacities is proposed, and their performance was verified. The difference between the proposed method and the other studies is that instead of a fixed load profile, the actual load profile changes instantaneously and randomly in the long run. Using national realistic data, the analysis is implemented and the results are compared with the existing system. Then, the benefits of the proposed method will be assessed for the improvement of Turkey interconnected system. To avoid overloading of power transmission lines and additional losses due to reactive power, loss minimization and voltage regulation may be provided using STATCOM (Static Synchronous Compensators) technology in a faster and more stable way. Reactive compensation is made with the help of static controllers and power electronics elements to increase the capacity, controllability of the power transferred by the power transmission lines and to ensure the reactive power demand of the system rapidly. Shunt reactive compensator devices can be designed with switching type converters based on semiconductor devices. FACTS (Flexible Alternating Current Transmission Systems) devices can produce and consume reactive power using switched converter circuits without the need for capacitor or reactor groups in the compensation of transmission lines. The use and development of FACTS (Flexible Alternating Current Transmission Systems) in power transmission systems bring many applications to improve the stability of power systems [1]. They are also used to increase the stability of the system and control the power flow. The greatest advantage of such devices is their flexibility and controllability [2]. These applications can be done by controlling the voltage value and phase angle [3]. The simulation study shows that STATCOM (Static Synchronous Compensators) responds very quickly to unexpected sudden voltage changes [4]. A methodology for introducing FACTS (Flexible Alternating Current Transmission Systems) models into the energy transmission network is described. STATCOM (Static Synchronous Compensators) application methodology for the application of power systems stability program to the power system

analysis program includes four stages. The results showed that the models in the power system analysis software were applied correctly and how FACTS (Flexible Alternating Current Transmission Systems) devices contributed to improving power system stability [5]. In the literature, studies on FACTS (Flexible Alternating Current Transmission Systems) technology have been carried out, and although FACTS (Flexible Alternating Current Transmission Systems) controllers have been classified and examined, a feasibility study has not been carried out using a FACTS (Flexible Alternating Current Transmission Systems) technology for a real power system model that has been constructed with realistic values taking into account all the parameters for a real energy transmission system.

II. METHOD AND MODELS

A. Load Flow Analysis

In the modeling of the power system of the pilot region where the feasibility study will be performed, load flow analysis is performed:

With the help of the node admittance matrix, the phasor current and voltage relations of the network can be written as matrices as follows.

$$\begin{bmatrix} \tilde{I}_1 \\ \tilde{I}_2 \\ \vdots \\ \tilde{I}_n \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1n} \\ Y_{21} & Y_{22} & \cdots & Y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nn} \end{bmatrix} \begin{bmatrix} \tilde{V}_1 \\ \tilde{V}_2 \\ \vdots \\ \tilde{V}_n \end{bmatrix} \quad (1)$$

Where, n : Total number of nodes

Y_{ii} : Self admittance of node i (sum of all admittances terminating in node i)

Y_{ij} : Common admittance between nodes i and j (inverse of the sum of all admittances between nodes i and j)

V_i : phasor voltage relative to earth at node i

I_i : Phasor current supplied to grid in node i

The effects of generators, non-linear loads, and other devices connected to the grid nodes are reflected in the node current. Constant impedance (linear) loads are also included in the node admittance matrix. In the formation of node equations, in non-linear power flow equations, the equation will be linear if I current inputs are known. The current inputs depend on the P, Q and V values in any k node:

$$\tilde{I}_k = \frac{P_k - jQ_k}{\tilde{V}_k} \quad (2)$$

$$\begin{bmatrix} \Delta P_2^{(k)} \\ \vdots \\ \Delta P_n^{(k)} \\ \vdots \\ \Delta Q_2^{(k)} \\ \vdots \\ \Delta Q_n^{(k)} \end{bmatrix} = \begin{bmatrix} \left(\frac{\partial P_2}{\partial \delta_2}\right)^{(k)} & \cdots & \left(\frac{\partial P_2}{\partial \delta_n}\right)^{(k)} & \left(\frac{\partial P_2}{\partial |V|_2}\right)^{(k)} & \cdots & \left(\frac{\partial P_2}{\partial |V|_n}\right)^{(k)} & \Delta \delta_2^{(k)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \left(\frac{\partial P_n}{\partial \delta_2}\right)^{(k)} & \cdots & \left(\frac{\partial P_n}{\partial \delta_n}\right)^{(k)} & \left(\frac{\partial P_n}{\partial |V|_2}\right)^{(k)} & \cdots & \left(\frac{\partial P_n}{\partial |V|_n}\right)^{(k)} & \Delta \delta_n^{(k)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \left(\frac{\partial Q_2}{\partial \delta_2}\right)^{(k)} & \cdots & \left(\frac{\partial Q_2}{\partial \delta_n}\right)^{(k)} & \left(\frac{\partial Q_2}{\partial |V|_2}\right)^{(k)} & \cdots & \left(\frac{\partial Q_2}{\partial |V|_n}\right)^{(k)} & \Delta |V|_2^{(k)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \left(\frac{\partial Q_n}{\partial \delta_2}\right)^{(k)} & \cdots & \left(\frac{\partial Q_n}{\partial \delta_n}\right)^{(k)} & \left(\frac{\partial Q_n}{\partial |V|_2}\right)^{(k)} & \cdots & \left(\frac{\partial Q_n}{\partial |V|_n}\right)^{(k)} & \Delta |V|_n^{(k)} \end{bmatrix}$$

$$\begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix} = \begin{bmatrix} J_1 & J_2 \\ J_3 & J_4 \end{bmatrix} \begin{bmatrix} \Delta \delta \\ \Delta |V| \end{bmatrix} \quad (4)$$

P and Q are specified for load busbars and P and V for voltage-controlled busbars. For other node types, the relations between P, Q, V and I are defined by the characteristics of the devices connected to those nodes. The problem of boundary conditions brought about by various types of nodes is turned into a nonlinear problem and solved iteratively using Fast-Decoupled Newton-Raphson Method. Fast-Decoupled Newton-Raphson Method, Newton-Raphson method in the Jacobian matrix phase angle dependence of the reactive power (J3) and active power to the voltage dependence (J2) is based on the principle of making load flow analysis is a method that is much faster accelerated. The equation systems obtained are given in equation (5) and (6).

$$J_1 = (i)\Delta\delta(i) = \Delta P(i) \quad (5)$$

$$J_4 = (i)\Delta V(i) = \Delta Q(i) \quad (6)$$

To further reduce the calculation time of the Fast-Decoupled method, the Jacobian matrix will be created according to the initial conditions and the Fast-Decoupled method with constant Jacobian will be applied during the calculation.

$$\begin{aligned} \frac{\partial P_i}{\partial V_j} &= 2V_i G_{ii} + \sum_{k=1, k \neq i}^n V_k Y_{ik} \cos(\theta_i - \theta_k - \alpha_{ik}) \\ &= 2V_i G_{ii} + \sum_{k=1, k \neq i}^n V_k Y_{ik} [\cos(\theta_i - \theta_k) \cos \alpha_{ik} + \\ &\quad \sin(\theta_i - \theta_k) \sin \alpha_{ik}] \\ &= 2V_i G_{ii} + \sum_{k=1, k \neq i}^n V_k [G_{ik} \cos(\theta_i - \theta_k) + B_{ik} \sin(\theta_i - \theta_k)]; \quad j = i \end{aligned} \quad (7)$$

$$\begin{aligned} \frac{\partial P_i}{\partial V_j} &= V_i Y_{ij} \cos(\theta_i - \theta_j - \alpha_{ij}) \\ &= V_i Y_{ij} [\cos(\theta_i - \theta_j) \cos \alpha_{ij} + \sin(\theta_i - \theta_j) \sin \alpha_{ij}] \end{aligned} \quad (9)$$

$$= V_i [G_{ij} \cos(\theta_i - \theta_j) + B_{ij} \sin(\theta_i - \theta_j)]; \quad j \neq i \quad (10)$$

$$\frac{\partial P_i}{\partial V_i} \approx 0 \quad \text{ve} \quad \frac{\partial P_i}{\partial V_j} \approx 0 \quad \Rightarrow \quad J_2 \approx 0 \quad (11)$$

$$\frac{\partial Q_i}{\partial \theta_j} = \sum_{k=1, k \neq i}^n V_i V_k [G_{ik} \cos(\theta_i - \theta_k) + B_{ik} \sin(\theta_i - \theta_k)]; \quad (12)$$

$$\frac{\partial Q_i}{\partial \theta_j} = -V_i V_j [G_{ij} \cos(\theta_i - \theta_j) + B_{ij} \sin(\theta_i - \theta_j)]; \quad j \neq i \quad (13)$$

$$\frac{\partial Q_i}{\partial \theta_i} \approx 0 \quad \text{ve} \quad \frac{\partial Q_i}{\partial \theta_j} \approx 0 \quad \Rightarrow \quad J_3 \approx 0 \quad (14)$$

$$\begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix} = \begin{bmatrix} J_1 & 0 \\ 0 & J_4 \end{bmatrix} \begin{bmatrix} \Delta \theta \\ \Delta V \end{bmatrix} \quad (15)$$

In this study, dynamic load flow analysis will be performed under discontinuously distributed generation and variable power demand situations and hourly changes of electrical parameters of busbars will be calculated. Thus, the effects of distributed generation, which is formed from sources showing production discontinuity, on the voltage and power factor stability of busbars can be analyzed in the face of changing power demands.

B. Static Synchronous Compensators Models

Because FACTS devices, which constitute modern compensation methods, react in a short time, control each phase separately, and compensate unbalanced loads, the use of these devices gains importance [6]. Since FACTS control is based on power electronics, they react faster than conventional controllers. These devices increase the stability limits of the transmission lines when used appropriately. FACTS have two primary purposes. The first one is to increase the power carrying capacity of the transmission systems; the second is to control the power flow over the transmission lines [7]. Today, many power flow controllers have been developed under the name FACTS. The most commonly used ones are; Static Yes Compensator (SVC), Thyristor Controlled Series Capacitor (TCSC), Static Compensator (STATCOM), Combined Power Flow Controller (UPFC), Phase Shifter and Static Synchronous Serial Capacitor (SSSC). STATCOM, known as Advanced Static Var Compensator (ASVC), is a FACTS controller, which is controlled to draw reactive current from the power system and connected to an inverter between a dc energy storage element and a three-phase system. The shunt is connected to the STATCOM transmission line. STATCOM is to regulate the voltage of the transmission line at the connection point by drawing a controlled reactive current from the transmission line. This process is the primary function of STATCOM [8]. In the simplest case, as shown in Figure 1, a STATCOM controller; It consists of a connection transformer, voltage source inverter, and DC energy storage element. Since the energy storage element is a tiny capacitor, it can only exchange reactive power with the STATCOM transmission system. The amplitude of the current flowing from the voltage source inverter to the line can be calculated by the following equation (16). Where X is the leakage reactance of the connection transformer, the reactive power received and exchanged can be expressed as in equation (17).

$$I_{ac} = \frac{V_0 - V_{av}}{X} \quad (16)$$

$$Q = \frac{V_0^2 - V_0 V_{ac} \cos \alpha}{X} \quad (17)$$

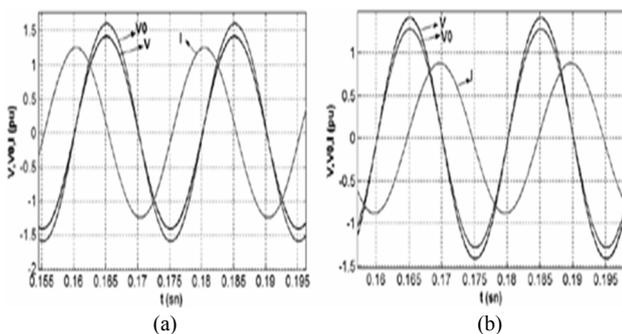


Fig.1 STATCOM Capacitive mode (a) and inductive mode (b)

By varying the amplitude of the 3-phase output voltage of the voltage-driven inverter, STATCOM can be controlled to generate or draw reactive power. If the output voltage (V_0) of the inverter is greater than the ac system voltage (V_{ac}), then the ac current (I_{ac}) flows from the inverter to the ac system generating reactive power via the transformer reactance. In this case, the inverter generates capacitive current for the ac system at an angle beyond its voltage. If the amplitude of the

inverter output voltage is smaller than the ac system voltage, the ac current flows from the ac system to the voltage source inverter. In this case, the inverter draws an inductive current at an angle behind the voltage, i.e, it consumes inductive reactive power. If the output voltage of the inverter and the amplitude of the ac system voltages are equal, there will be no ac current flow from the inverter to the ac system or from the ac system to the inverter. In short, the inverter will not produce or consume reactive power [9]. The active power exchange between the voltage source inverter and the AC system can be calculated using equation (18).

$$P = \frac{V_0 V_{ac} \sin \alpha}{X} \quad (18)$$

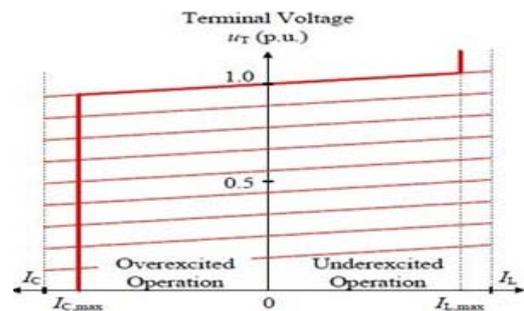


Fig 2. STATCOM, V-I characteristics

Figure 2 shows the STATCOM V-I characteristics. At low voltages, STATCOM's current supply capacity is much better. STATCOM can provide either full capacitive or full inductive output current at any system voltage. The amount of reactive power compensation provided by STATCOM is more significant because, at a low voltage level, reactive power decreases proportionally to the square of the voltage, while STATCOM decreases linearly with the energy. This makes STATCOM's reactive power much more controllable [10].

III. SIMULATION STUDY

The purpose of the load flow analysis is to examine power flows, loads, busbar voltages in possible variable load situations. As a result of this analysis, many issues such as voltage increases in busbars, loadings in lines and transformers, power flows that can change direction, reactive power capacity of the production plant can be monitored. As described in the above section, STATCOM devices are active devices used in voltage control within their limits. For example, it was modeled on ten busbar systems, and their effectiveness was investigated. Since the STATCOM devices are connected in parallel to the electrical power system, no auxiliary virtual bus is needed. In the modeling, the upper and lower limit values are modeled as susceptance and can work with both inductive and capacitive characteristics. STATCOM device is modeled to busbar Denizli in the system. The ten-bar power transmission system using the graphical interface of the program is shown in (Figure 10).

A. Power System Modeling

If proper and adequate equipment is not used correctly in power transmission, the power system becomes weak against steady-state and transient problems, and stability limits will change. To test the accuracy and performance of the proposed method, a realistic Energy Transmission System network model consisting of ten bus was formed in the energy transmission system. In the modeled power system, data for the network: 160 kV voltage level (1 pu), voltage lower range (0.9-1pu) upper voltage range (1-1.1pu), total 600 MVA (max) YNd5 transformer power, 500 MW (G1-G2-G3) full generation capacity, 400 MW (load ABCDE) whole load, Short circuit power (Sk) 10000 MVA with external network and voltage controlled 20 MVAR (C) Capacitor connected.

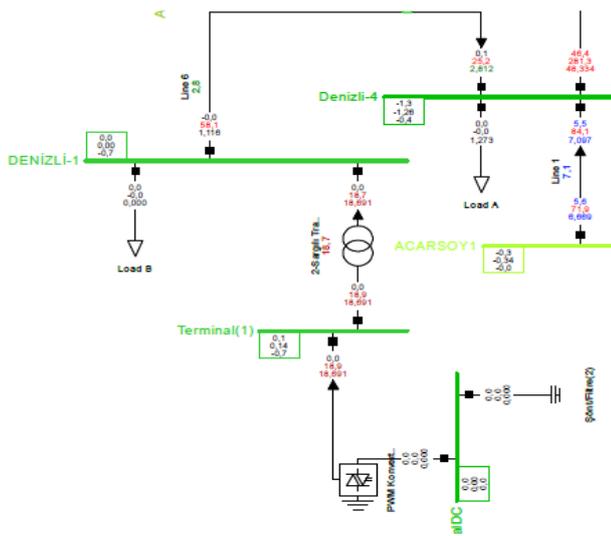


Fig.3 Busbar-Denizli STATCOM connection

The simulation study was carried out with balanced 3-phase system voltages during 4-second steps and a total of 20-second operation. Modeled power transmission system line and bus loadings are between 15%, and 20% voltage 1 pu levels, system production G1 9,12%, G2 68,41%, and G3 74,39%. Simulation study has been carried out for continuous or transient state faults whether 20 MVAR capacitors or STATCOM device is connected. Simulation studies have been carried out for load failure situations due to the failure of 30MW D load connected to Line-2. In both cases, the voltage values of the energy transmission system network model for other busbars were examined.

B. Results

The identification, development, and classification of FACTS controllers are discussed, and the realization of the results obtained by performing a simulation study by modeling the actual values for the Western Mediterranean Region electric power transmission system and determining the benefits provided are determined by realistic parameters. STATCOM, which is one of the FACTS controllers, is used for power system modeling, analysis and simulation programs, modeling and analysis for various conditions such as voltage drop, frequency changes, instantaneous energy generation and to improve the results obtained, system

reliability and stability, failure and hardware failure. Limiting the effects of replications to avoid power failure and Turkey interconnected by increasing the balance of the voltage and power control of the system more efficient, have reviewed the contribution on the quality and economical way of providing electrical energy continuum. The electrical energy transmission system of the Western Mediterranean region was modeled on ten busbar systems in Denizli region as an example, and their effectiveness was investigated. Since STATCOM devices are connected in parallel to the electrical power system, no auxiliary virtual bar is needed. In the modeling, lower and upper limit values are modeled as 160 kV voltage level (1 pu), voltage lower range (0.9-1pu) and high voltage range (1-1.1 pu) and can work both inductive and capacitive characteristics. STATCOM devices are modeled separately for Denizli-1 busbar. In case of failure of one of the lines (line-2a) in the electric power transmission system of the shaped Western Mediterranean region; The effects of the FACTS devices on the stability limits of the STATCOM power system against voltage collapses were investigated, and the impact against voltage oscillations in the order after the disturbance effects were evaluated. (Figure-4 & Figure-5)

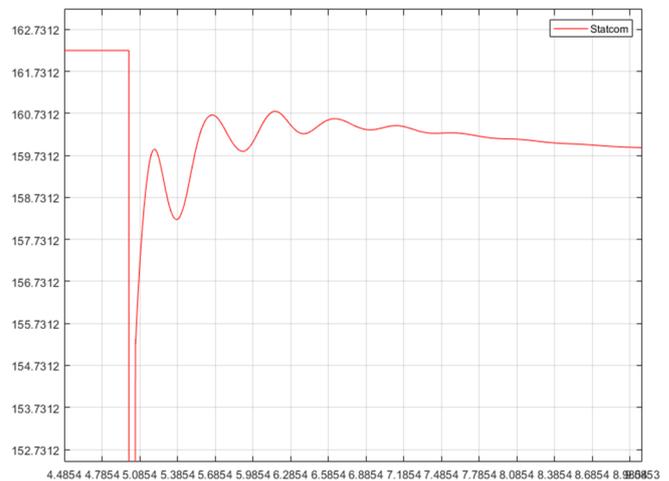


Figure 4. Stability limits against STATCOM voltage failures

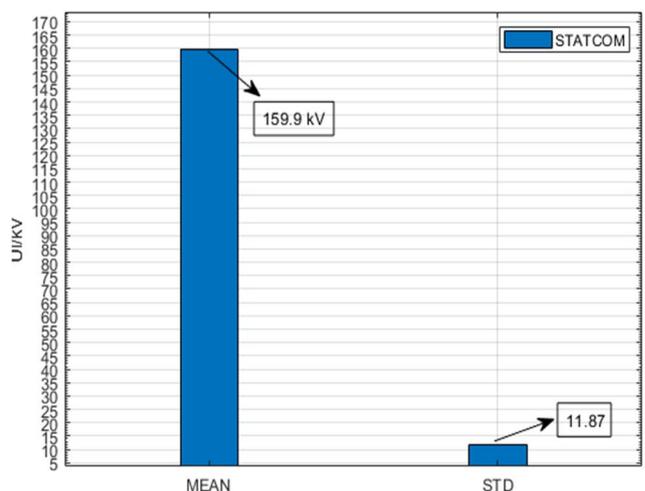


Fig 5. STATCOM mean and deflection values against voltage failures

By evaluating the voltage-loading parameter curves for the most critical busbars experiencing voltage problems in the modeled electric power transmission system of the Western Mediterranean region, the contribution of the STATCOM

systems to the reactive power is examined by connecting STATCOM systems, and in case the lines supplying the critical busbar of the system are disabled, STATCOM is loaded. And the steady-state power flow analysis has been done, and in case there are different disturbing effects in the system, the voltage stability change has been shown by adding STATCOM to Denizli-1 busbar as a result of the power flow. (Figure 6)

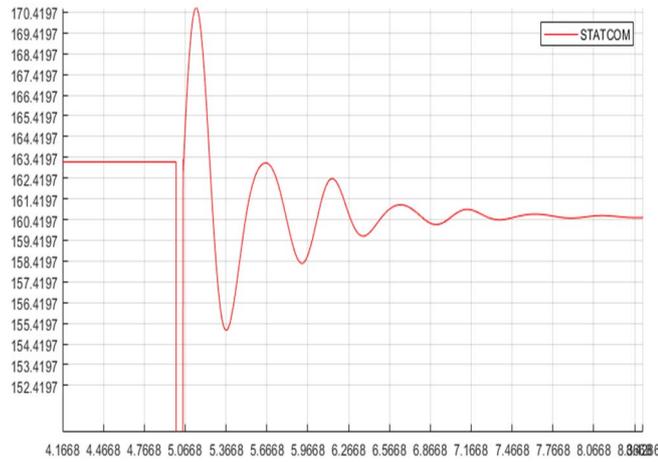


Fig 6. Load limits of critical bus to STATCOM voltage failures

Modeling the West Mediterranean region electricity energy transmission system and performing load flow analysis as a result of STATCOM voltage regulation added parallel to the busbars needed, the points where the system undergoes voltage collapse were determined, and the effectiveness of FACTS devices at this point were compared. Increased. STATCOM's reactive power load value of the system works most stable, active and reactive power values provide the best power transfer, STATCOM required reactive power compensation is produced quickly, voltage collapse busbar voltage value sudden load changes and despite the desired reference value It has been seen. When the rotor angle and rotor speed oscillations of the generators are examined, it has been shown that the contributions made to the oscillation suppression improve the generator voltage stability and help to increase the reliability and capacity of the system, thus providing the system with more inductive and capacitive energy (Figure-7, Figure-8).

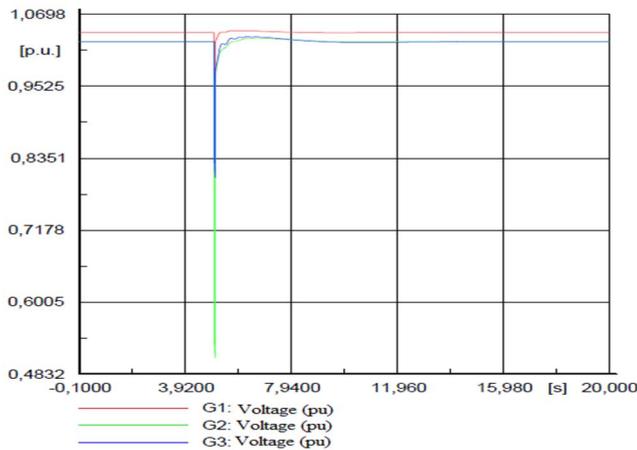


Fig 7. STATCOM generator voltage stability

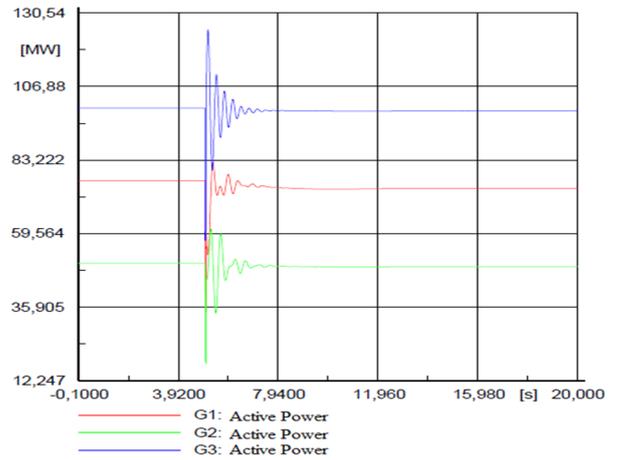


Fig 8. Power oscillations of STATCOM generators

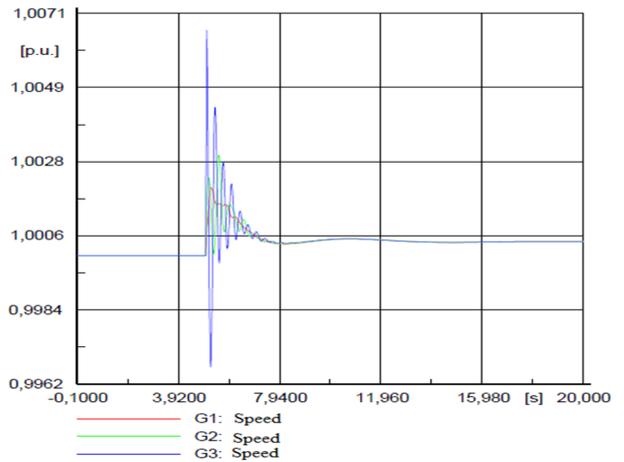


Fig 9. Rotor angle and rotor speed oscillations of STATCOM generators

The bus voltages are presented for the pre-fault and after the fault with the usage of Capacitor Bank and STATCOM device (Table I & II). As it can be seen from the results, STATCOM device regulates the voltage much better compared to capacitor banks. The bus voltages almost were not affected from the fault by the STATCOM usage (Table II).

TABLE I.
PRE-FAULT STATCOM BUS VOLTAGES

Bus kV/(pu)	PRE-FAULT BUS VOLTAGE	
	Capacitor Enabled	Statcom Enabled
DENİZLİ-1 kV/(pu)	161,9342 1,01208	160,0 1,0
DENİZLİ-2 kV/(pu)	163,2398 1,02024	163,1085 1,01942
DENİZLİ-3 kV/(pu)	158,4197 0,99012	158,3302 0,98956
DENİZLİ-4 kV/(pu)	161,4339 1,008961	161,0137 1,00633
ACARSOY kV/(pu)	165,177 1,03235	165,08 1,03175
NETWORK kV/(pu)	160,8582 1,005361	160,7506 1,00469

TABLE II.
STATCOM AND BUS VOLTAGES AFTER FAILURE

Bus kV/(pu)	BUS FAILURES AFTER FAILURE	
	Capacitor Enabled	Statcom Enabled
DENİZLİ-1 kV/(pu)	157,8052 0,98628	160,0 1,0
DENİZLİ-2 kV/(pu)	162,1797 1,01362	162,3029 1,01439
DENİZLİ-3 kV/(pu)	157,8219 0,98638	157,9073 0,98692
DENİZLİ-4 kV/(pu)	158,2907 0,98931	158,8408 0,99275
ACARSOY kV/(pu)	164,3486 1,02717	164,4754 1,02797
NETWORK kV/(pu)	159,9385 0,99961	160,0793 1,00049

IV. CONCLUSION

In this study, the effects of STATCOM controllers on power system voltage stability are investigated. As a result of STATCOM usage, the voltage stability of the model network was observed. Also, it has been observed on the ten-bus power system how the voltage instability occurs at the different load characteristics.

Stability values of transmission lines and voltage stability limit values of load bus was calculated. After connecting STATCOM devices, load flow studies were performed by using power system analysis program.

Voltage regulation may be provided by inserting STATCOM devices into the Turkey's interconnected system, and the losses of the power transmission line may be reduced and electricity energy may be transmitted and distributed in more secure and controllable way. By this way, significant economic developments may be ensured.

Implementing the feasibility study by using STATCOM device, power transmission costs are aimed to be reduced technically, by improving the sensitivity of voltage control to stabilize the production and consumption balance.

When the system is evaluated in terms of oscillating operation and voltage drop, it is observed that the use of parallel and serial connected compensators together makes the system more stable than single-use. FACTS devices are able to raise the highest load point of the system against voltage collapse higher than the manual systems (current situation). In case of the failure of one of the lines in the

network, bus voltages, generator rotor speeds and rotor angles oscillating operation mode can be suppressed with FACTS devices. STATCOM power system increases the maximum load limits. Furthermore, when the disturbance of the power system (line failure) occurs, the reactive power compensation may be implemented with the proposed system and voltage collapse is prevented under overload conditions. It is observed that the proposed system removes voltage oscillations in the generation busbars in case of the short circuit in the system.

ACKNOWLEDGMENT

This study has been supported by Akdeniz University Scientific Research Projects Coordination Unit within the scope of the project FBA-2018-3792.

REFERENCES

- [1] Hingorani N. G. "Flexible AC Transmission" IEEE reprinted from IEEE Spectrum, Vol. 30, No.4 1993 pp 40-45.
- [2] Cheng, H. In, I. and Chen S. "DC-Link Voltage Control and Performance Analysis of STATCOM" 2002.
- [3] Yang, Z. Shen, C. Zhang, L. Crow M. L." Integration of a STATCOM and Battery Energy Storage", IEEE Trans. on Power System, Vol. 16, no. 2, May 2001, pp. 254-260.
- [4] Çötel, "STATCOM ile güç akış kontrolü", Yüksek Lisans Tezi Elektrik Eğitimi, Fırat Üniversitesi Fen Bilimleri Enstitüsü Elazığ, 2006.
- [5] Jaime Cepeda, Esteban Agüero, "FACTS models for stability studies in DİGSILENT Power Factory" IEEE Transmission and Distribution Latin America 2014, DOI: 10.1109/TDC-LA.2014.6955182.
- [6] TMMOB, "Reaktif Güç Kompanzasyonu Seminer Notları" İstanbul EMO, 1999.
- [7] A. Hasanovic, "Modeling and Control of The Unified Power Flow Controller (UPFC)", MA Thesis, West Virginia Uni., 2000.
- [8] Schauder, C. and Mehta, H. "Vector Analysis and Control of Advanced Static VAR Compensators", IEE Proceedings-C, Vol. 140, No. 4, 1993 pp.299-306.
- [9] Gyugyi, L. "Power Electronics in Electric Utilities: Static Var Compensators", Proceedings of The IEEE, vol. 76, no. 4, 1988. pp. 483-493.
- [10] Paserba J., (2003), How FACTS Controllers Benefit AC Transmission System, Trans. And Dist. Con. and Exp., IEEE PES Volume 3, 7-12 Sept. vol.3 949 – 956.
- [11] H. Feza Carlak, E. Kayar, "Voltage Regulation Analysis in Energy Transmission Systems Using TCR-TSC SVC Systems" 4th International Mediterranean Science and Engineering Congress IMSEC Alanya 2019 pp. 476-481.

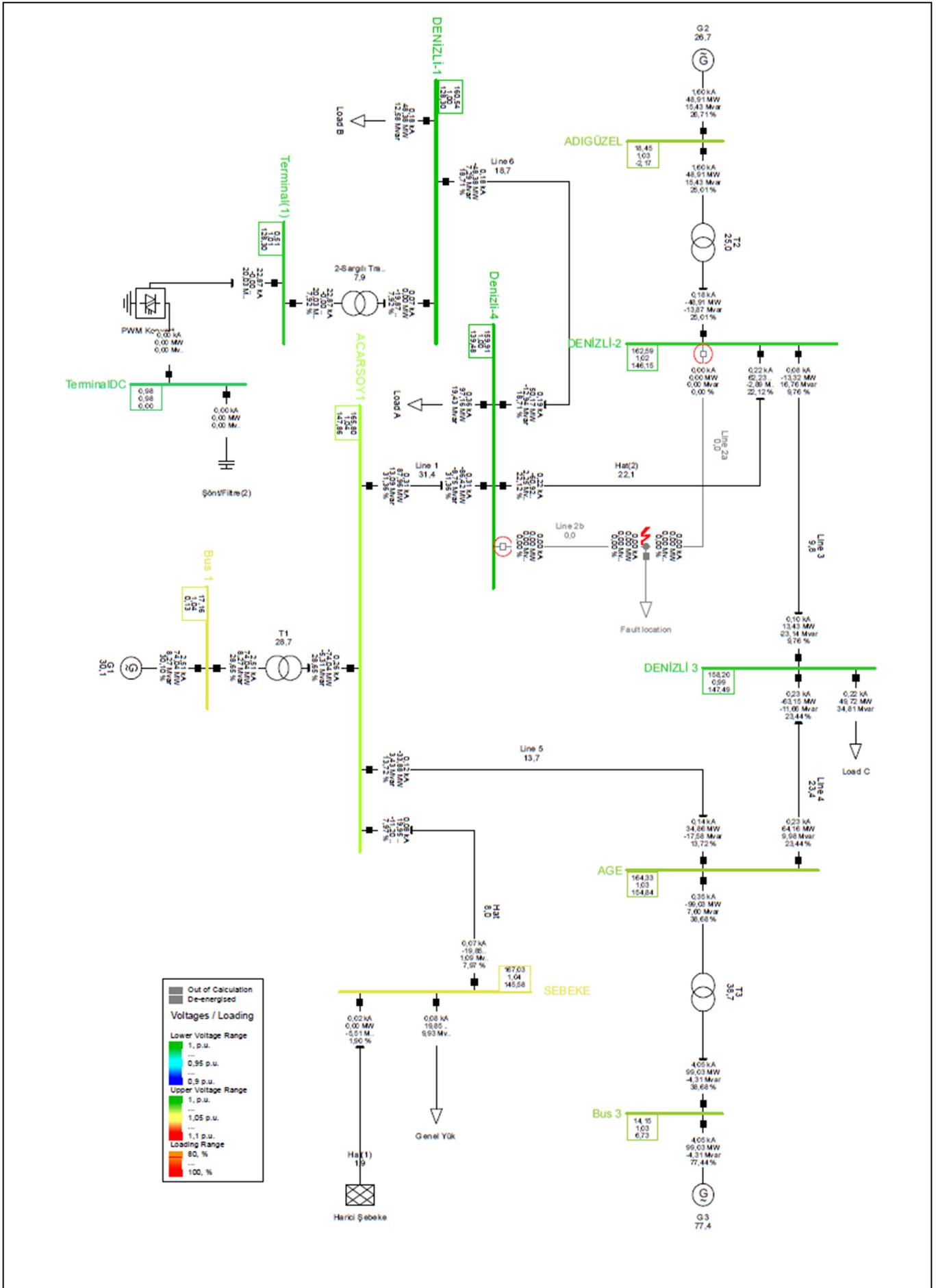


Fig -10 Western Mediterranean / Denizli Energy Transmission System Network Model

Research of the effect of discrete light sources on seeds of vegetable and green cultures and the possibility of their approximation to modified sunlight

Danila Yu. Donskoy
Don State Technical University
Rostov-on-Don, Russia
dand22@bk.ru

Alexander D. Lukyanov
Don State Technical University
Rostov-on-Don, Russia
alexlukjanov1998@gmail.com

Eugenia P. Kluchka
Don State Technical University
Rostov-on-Don, Russia
klyuchkae@mail.ru

Marko Petković
University of Kragujevac
Čačak, Serbia
marko.petkovic@kg.ac.rs

Abstract — This article discusses the method of increasing the germination and speed of seed development using the optimal ratio of spectral components of intensity and photoperiod required for the seeds of vegetable and green cultures at the germination stage. The method is based on the approximation of the limited spectra of discrete sources of illumination to a continuous modified spectrum of sunlight by the method of the least squares in order to minimize the square of the approximation error. Six different types of LEDs (red, green, blue, white (RGB), “full-spectrum” (RGBWFI) and infrared) was used for the approximation of the continuous spectrum for radish seeds. The experiment period was three days. After the three days of experimental seedlings, the average root length (stimulated by RGBWFI LED) was 9.4 mm, the total length of the germ 24.7 mm, the root thickness 0.85 mm and the total germination percentage 89.7%. We research is being conducted on various seeds to accurately optimize the method for a specific seed culture. The found efficiency criteria will provide an opportunity to systematize numerous well-known results of the impact of variable light modes in order to identify plants as a biological object by optical properties.

Keywords — approximation, spectrum, LEDs, light stimulation, the method of the least squares, radish

I. INTRODUCTION

Currently, when growing seedlings, smart lamps with spectrum optimization are practically not used. First of all, this is due to the lack of valid methods that allow optimizing the spectrum of illumination at various stages of plant development. Popular “full-spectrum” LED matrices, with a predominance of the red and blue components of the spectrum, are optimized “by average” and do not take into account the characteristics and needs of a particular culture at different stages of plant development.

Existing studies are focused and controversial, it is known [1,2] that the entire spectral range of sunlight is important for plant development, and limiting the spectrum can lead to growth retardation and degradation.

At the same time, at different stages of plant development, a different spectral composition of illumination is optimal, which opens up opportunities both for growth light stimulation and for increasing the energy efficiency of

cultivation [3, 4, 5]. However, it remains an open question about the technological and algorithmic possibility of approximation of light of a given spectral composition by a set of sources with limited spectrum, as the basis for conducting research in this direction.

In this paper, the authors, on the basis of previously performed studies [6, 7, 8], present a technique for approximating a continuous spectrum by a set of limited spectra with minimization of the quadratic error over the range of the continuous spectrum. The technique is intended for use in intelligent lamps for agricultural cultivation, providing the optimal spectral composition and intensity of illumination at various stages of cultivation and for different cultures.

Thus, the aim of the work is to increase the effectiveness of agrotechnologies of a closed ground by providing the possibility of using promising agrophotonic algorithms that require control of the spectral composition, intensity and photoperiod of illumination.

To achieve this goal, the authors solved the following tasks:

- Developed a method for determining the coefficients for the approximation of the continuous spectrum by a set of limited spectra based on the method of least squares.
- The technique was tested on a number of continuous spectra corresponding to the penetration of sunlight into the soil at different depths
- On the basis of certain coefficients, the law of control of an intelligent agroform was implemented to stimulate seed germination and its positive effect on the indicators of their development was experimentally shown.

The presented technique is new both in its content and in the field of its implementation in intelligent lamps of agrotechnological purpose.

II. CREATING A PHYTO-ILLUMINATION ALGORITHM WITH A SPECIFIED RATIO OF SPECTRAL COMPONENTS

The general mathematical solution of this problem using the method of the least squares was considered by us in the article "Simulation, spectrum synthesis". There are other studies in which the Gauss method was used for approximation, but our method is more versatile and adapted to a wide range of discrete light sources [9]. The problem was solved by using the system of intellectual preparation of seeds for germination. For the approximation, six types of LEDs were used: red, green, blue, white (RGBW), "full-spectrum" (F), having high intensity in the blue and red zones of the visible light spectrum, and infrared (IR).

The spectral characteristics of discrete light sources can be seen in Fig. 1. (SPF - spectral photon flux.)

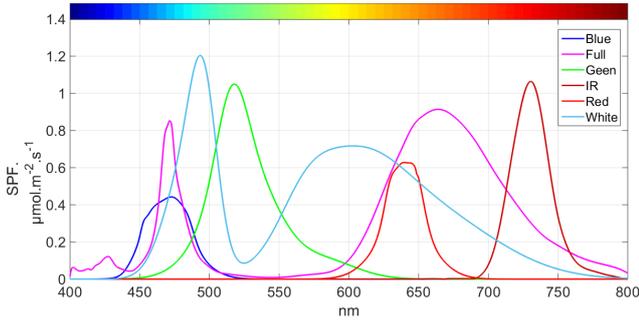


Fig. 1. Spectral characteristics of RGB WFI LEDs.

The solution to this problem is represented by J (kn) (1).

$$J(k_1, k_2, \dots, k_n) = \int_{\lambda_1}^{\lambda_2} \left(S(\lambda) - \sum_{i=1}^n k_i S_i(\lambda) \right)^2 d\lambda \xrightarrow{k_1, \dots, k_n} \min \quad (1)$$

From here we obtain the expression (2), where k is the desired approximation coefficients.

$$\frac{J}{\partial k_i} = \int_{\lambda_1}^{\lambda_2} S(\lambda) S_i(\lambda) d\lambda + \sum_{j=1}^n k_j \int_{\lambda_1}^{\lambda_2} S_i(\lambda) S_j(\lambda) d\lambda \quad (2)$$

Let's name spectral densities of spectra of discrete light sources $S_{red}(S_1)$, $S_{green}(S_2)$, $S_{blue}(S_3)$, $S_{white}(S_4)$, $S_{full}(S_5)$, $S_{ir}(S_6)$. The spectral density of the modified spectrum of sunlight on the surface and at a depth of 3, 6, 9 mm denotes S_g , S_{g3} , S_{g6} , S_{g9} . Spectral range - λ 400 to 800 nm. For approximation, it is necessary to find the following integral components of the matrix of the system of equations A (3) and B (4):

$$A = \begin{bmatrix} \int_{\lambda_1}^{\lambda_2} S_1(\lambda) S_1(\lambda) d\lambda & \dots & \int_{\lambda_1}^{\lambda_2} S_1(\lambda) S_6(\lambda) d\lambda \\ \vdots & \ddots & \vdots \\ \int_{\lambda_1}^{\lambda_2} S_6(\lambda) S_1(\lambda) d\lambda & \dots & \int_{\lambda_1}^{\lambda_2} S_6(\lambda) S_6(\lambda) d\lambda \end{bmatrix} \quad (3)$$

$$B = \begin{bmatrix} \int_{\lambda_1}^{\lambda_2} S_{(g, g3, g6, g9)}(\lambda) S_1(\lambda) d\lambda \\ \vdots \\ \int_{\lambda_1}^{\lambda_2} S_{(g, g3, g6, g9)}(\lambda) S_6(\lambda) d\lambda \end{bmatrix} \quad (4)$$

From here you can find the approximation coefficients for each spectrum $k_1, 2, 3, 4, 5, 6$ using the following (5):

$$k_{1,2,3,4,5,6} = (A^T A)^{-1} A^T B \quad (5)$$

As a result, the approximation coefficients: $k_1=0.3992$, $k_2=1.0815$, $k_3=0.0859$, $k_4=0.9603$, $k_5=1.1275$, $k_6=0.777$.

Graphs of comparison of the obtained spectra and modified spectra of sunlight are given below Fig. 2 a, b, c and Fig. 3. To reduce the intensity of the spectra at a length of 400 to 550 nm, long-wavelength filters with a transmittance of 550 nm Fig. 2 b, c and Fig. 3.

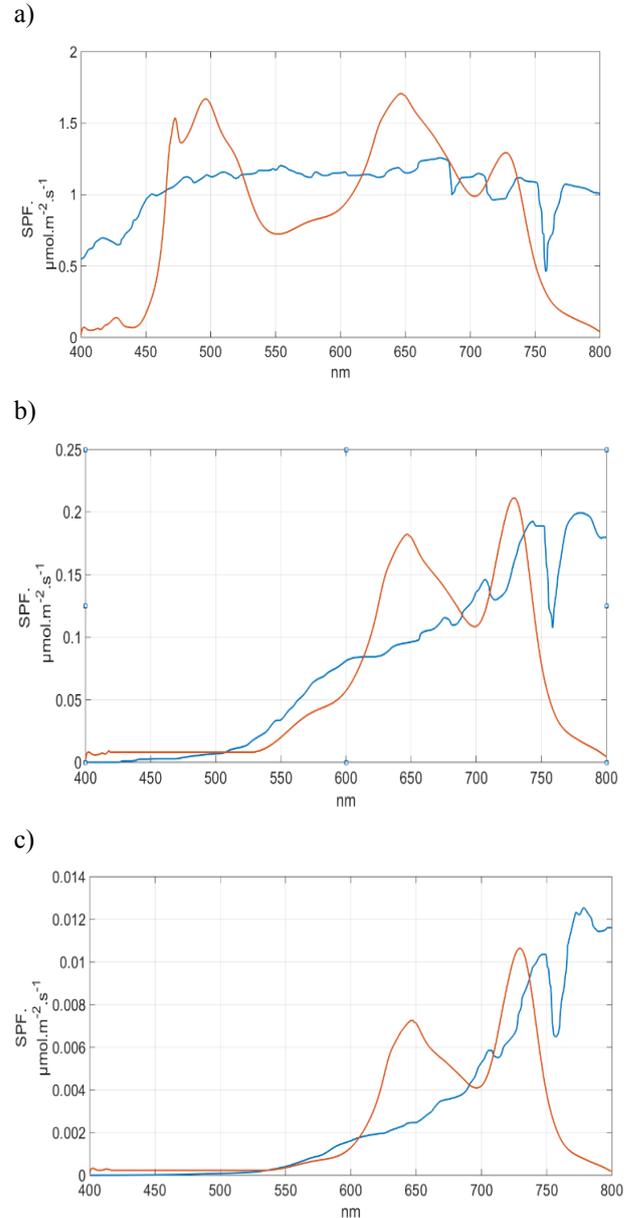


Fig. 2. Comparison of approximated spectra with the modified spectrum of sunlight a) surface, b) 3 mm, c) 6 mm.

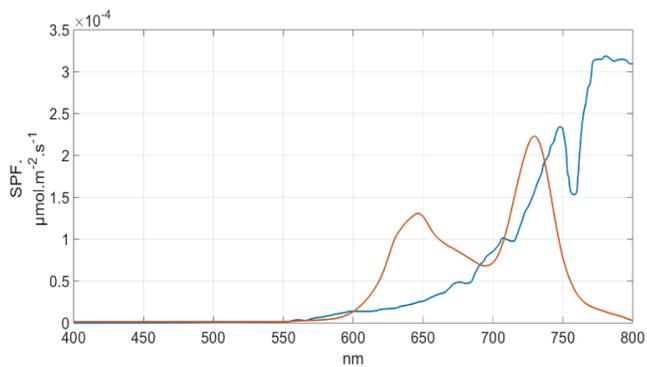


Fig. 3. Comparison of approximated spectra with the modified spectrum of sunlight 9 mm.

Studies have also been conducted on the approximation of simpler LED matrices, which are based on the RGBW or RGB full spectra Fig.3.

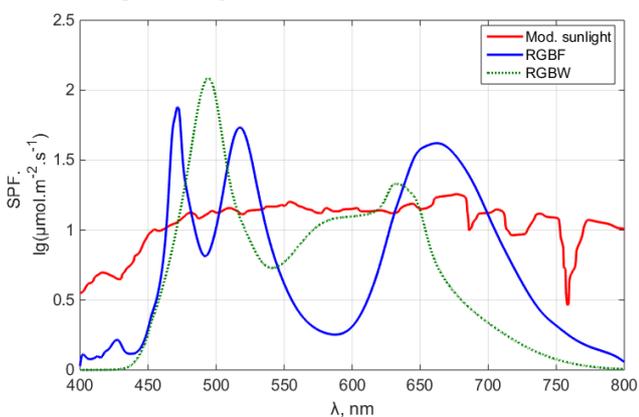


Fig. 4. Comparison of approximated spectra of discrete light sources to modified sunlight on the soil surface [7].

These indicators of the ratio of spectral components were applied in the algorithm of the system for preparing seeds for germination.

III. DESCRIPTION OF THE EXPERIMENTAL SETUP

For the formation of illumination with a given spectral characteristic and intensity, an experimental sample was developed that allows preparing seeds for germination and phenotyping (determining the stages of seed development by machine vision). The degree of development of the objects of the study was determined using the camera. Processing was carried out on raspberry PI 3.

The system allows to evaluate the criteria for the effectiveness of exposure to different light spectra, responds to the stage of seed development, adapting the characteristics of illumination Fig. 4 [8,9].

IV. THE DISCUSSION OF THE RESULTS

Experimental studies were carried out at constant: temperature of 27 C, substrate humidity of 100 %, illumination intensity from 0 to 5000 lux. The CO₂ concentration over the entire test period was in the range of 600-1100 ppm, there was also no direct air flow to the seed, which prevented weathering. In the role of the substrate was foam rubber ST 18/20 10 mm

development to the seedling stage is about 3-4 days. All seeds from one batch. On Fig. 5. depicts one of the obtained shoots.

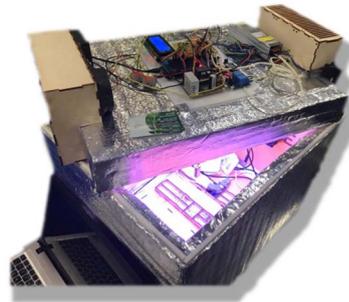


Fig. 5. System for experimental studies on the light stimulation of vegetable and green cultures.



Fig. 6. A batch of seeds and an example of a batch of radish sprout after exposure to the formed light, approximated to solar.

According to the results of these experiments (experiment period 3 days), we obtained seedlings with the following parameters: the average root length is 9.4 mm (1 - 3 mm more than with the RGBW light-stimulation spectra separately), the total length of the germ is on average 24.7 mm, the average maximum diameter is 0.85 mm, color from RGB (112, 131, 13) to RGB (166,162,63). The germination percentage is 89.7 %, which is on average 15.2 % higher than when using a separate spectral range of RGBW LEDs Fig. 6.

V. CONCLUSION

Based on the above, it can be argued that the tasks have been successfully implemented, the developed methodology has proved its effectiveness, and the goal of the research has been successfully achieved. The seed of vegetable and green cultures is a complex biological object that requires extensive research. Universal light-stimulation algorithms allow on average to increase the germination of some

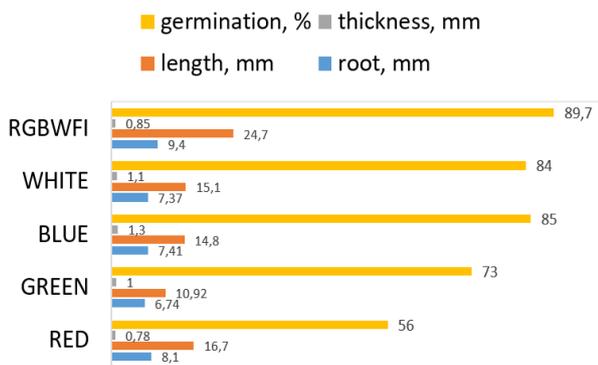


Fig. 7. Diagram of characteristics of the obtained experimental material as a result of light stimulation.

cultures, but for others, they may not be applicable. Research niches determine the methodology for preparing seeds of various cultures. Due to the formation of the optimal ratio of spectral components, the intensity of illumination and its periodicity, while maintaining favorable microclimate conditions in parallel, it is possible to significantly increase the productivity of existing greenhouses and develop light stimulation algorithms individually for different plants. Only due to the flexibility of control of discrete sources of illumination it is possible to determine the criteria for the effectiveness of certain methods of processing seeds of vegetable and green cultures [20]. The next stage of development will be conducting point studies to optimize this technique for certain types of seeds.

ACKNOWLEDGMENT

These studies were carried out in the framework of research, supported by the fund to promote innovation (the program "UMNIK") and the Don State Technical University.

REFERENCES

[1] Leyla Bayat, Mostafa Arab, Sasan Aliniaiefard, Mehdi Seif, Oksana Lastochkina, Tao Li, Effects of growth under different light spectra on the subsequent high light tolerance in rose plants, *AoB PLANTS*, Volume 10, Issue 5, October 2018, ply052, <https://doi.org/10.1093/aobpla/ply052>

[2] Kim, Hyeon-Hye & Wheeler, Ray & Sager, John & C Yorio, Neil & Goins, Gregory. (2005). Light-emitting diodes as an illumination source for plants: A review of research at Kennedy Space Center. *Habitation* (Elmsford, N.Y.). 10. 71-8. 10.3727/154296605774791232.

[3] A. Žukauskas, Z. Bliznikas, K. Breivė, G. Tamulaitis, G. Kurilėik, A. Novičkovas, P. Duchovskis, R. Ulinskaitė, A. Brazaitytė, J. Šikšnianienė. Semiconductor Lamp for Investigation and Control of Photophysiological Processes in Plants // *Electronics and Electrical Engineering*. – Kaunas: Technologija, 2004 – No.7(56). – P. 74- 79.

[4] Massa, Gioia D., Hyeon-Hye Kim, Raymond M. Wheeler, and Cary A. Mitchell. "Plant Productivity in Response to LED Lighting", *HortScience horts* 43, 7: 1951-1956, accessed Jul 10, 2019, <https://doi.org/10.21273/HORTSCI.43.7.1951>

[5] Kim K, Kook H, Jang J, Lee W, Kamala-Kannan S, et al. (2013) The Effect of Blue-light-emitting Diodes on Antioxidant Properties and Resistance to Botrytis cinerea in Tomato. *J Plant Pathol Microbe* 4: 203. doi:10.4172/2157-7471.1000203

[6] D. Yu. Donskoy, A. D. Lukyanov, M. A. Verzezi «Simulation, identification and dynamic control of the luminaire of the synthesized spectrum».- MATEC Web of Conferences [Электронный ресурс]. - 2018. - Vol. 226. - Номер статьи 02030. - (XIV International

Scientific-Technical Conference "Dynamic of Technical Systems" (DTS-2018); Rostov-on-Don, Russian Federation, September 12-14, 2018).- URL: <https://doi.org/10.1051/mateconf/201822602030>.

[7] Islam, Obaidul & Matsui, Shuichiro & Ichihashi, Syoichi. (1999). Effects of Light Quality on Seed Germination and Seedling Growth of *Cattleya Orchids* in vitro. *Engei Gakkai Zasshi*. 68. 1132-1138. 10.2503/jjshs.68.1132.

[8] Evgeniya P. Klyuchka1*, Viktor V. Radin1, Leonid M. Groshev1, Valeriy P. Maksimov2 Modelling a complex technical system of greenhouse production: the foundations of an interdisciplinary approach / MATEC Web of Conferences Volume 226, 2018 XIV International Scientific-Technical Conference «Dynamic of Technical Systems» (DTS-2018) <https://doi.org/10.1051/mateconf/201822602019>. Ссылка: https://www.mateconferences.org/articles/mateconf/abs/2018/85/mateconf_dts2018_02019/mateconf_dts2018_02019.html

[9] Evgeniya P. Klyuchka1*, Viktor V. Radin1, Leonid M. Groshev1, Sergey I. Kambulov2 Problems of modeling of complex technological systems of greenhouse production / MATEC Web of Conferences Volume 226, 2018 XIV International Scientific-Technical Conference «Dynamic of Technical Systems» (DTS-2018) <https://doi.org/10.1051/mateconf/201822602020>. Ссылка: https://www.mateconferences.org/articles/mateconf/abs/2018/85/mateconf_dts2018_02020/mateconf_dts2018_02020.html

[10] Feng Tian. Study and optimization of lighting systems for plant growth in a controlled environment. *Chemical and Process Engineering*. Université Paul Sabatier - Toulouse III, 2016. English. (NNT: 2016TOU30248). (tel-01582072): <https://tel.archives-ouvertes.fr/tel-01582072>

[11] D. BLISS, H. SMITH. Penetration of light into soil and its role in the control of seed germination. April 2006 *Plant Cell and Environment* 8(7):475 – 483. DOI10.1111/j.1365-3040.1985.tb01683.x

[12] Mark Tester, CHRISTINA MORRIS. The penetration of light into soil. April 2006 *Plant Cell and Environment* 10(4):281 – 286. DOI10.1111/j.1365-3040.1987.tb01607.x

[13] Donskoy D. Yu., Lukyanov A.D., Verzezi M.A., Katin O.I. "Development of automated systems for the intensification of the technological process of seed preparation by agrophotonic methods".- Modern informatization problems in the technological and telecommunication systems analysis and synthesis (MIP-2019'AS): Proceedings of the XXIV-th International Open Science Conference (Yelm, WA, USA, January 2019)/ Editor in Chief Dr. Sci., Prof. O.Ja. Kravets. - Yelm, WA, USA: Science Book Publishing House, 2019. – 5 p. (299-303)

[14] D. F. MANDOLI, G. A. FORD, L. J. WALDRON, J. A. NEMSON, W. R. BRIGGS. Some spectral properties of several soil types: implications for photomorphogenesis. *Plant, Cell & Environment*, Volume 13, Issue 3, April 1990, Pages 287–294, DOI: 10.1111/j.1365-3040.1990.tb01313.x

[15] A.V. Patsukov, A. P. Mishanov, S. A. Rakutko, A. E. Markov, V. N. Sudachenko "Influence of the optical radiation spectrum on the quality of tomato seedlings" *UDC*: 628.941.8: 581.14

[16] Vaniček, P. (1969). Approximate spectral analysis by least-squares fit - Successive spectral analysis. *Astrophysics and Space Science*. 4. 387-391. 10.1007/BF00651344.

[17] Joel Flores, Claudia González-Salvatierra, Enrique Jurado, Effect of light on seed germination and seedling shape of succulent species from Mexico, *Journal of Plant Ecology*, Volume 9, Issue 2, April 2016, Pages 174–179, <https://doi.org/10.1093/jpe/rtv046>

[18] J. Li, L.Y. Yin, M.A. Jongsma, C.Y. Wang, Effects of light, hydropriming and abiotic stress on seed germination, and shoot and root growth of pyrethrum (*Tanacetum cinerariifolium*), *Industrial Crops and Products*, Volume 34, Issue 3, 2011, Pages 1543-1549, ISSN 0926-6690, <https://doi.org/10.1016/j.indcrop.2011.05.012>.

[19] Joey H. Norikane. The Potential of LEDs in Plant-based Biopharmaceutical Production (2015) / *H ORT SCIENCE VOL. 50(9) S SEPTEMBER 2015*. – 1289-1292.

[20] Joey H. Norikane. The Potential of LEDs in Plant-based Biopharmaceutical Production (2015) / *H ORT SCIENCE VOL. 50(9) S SEPTEMBER 2015*. – 1289-1292.

Deriving adaptive homing sequences for weakly initialized nondeterministic FSMs

Evgenii Vinarskii
Computer science department
Lomonosov Moscow State
University
Moscow, Russia
vinevg2015@gmail.com

Aleksandr Tvardovskii
Radiophysical department
Tomsk State University
Tomsk, Russia
tvardal@mail.ru

Larisa Evtushenko
Computer science department
Higher School of Economics
Moscow, Russia
evtlarisa@mail.ru

Nina Yevtushenko
Software engineering
department
Ivannikov Institute for System
Programming
Moscow, Russia
nyevtush@gmail.com

Abstract—State identification sequences, such as homing and distinguishing sequences (HS and DS), are widely used in FSM (Finite State Machine) based testing in order to reduce the size of a returned complete test suite as well as minimize checking efforts in passive testing. Preset HS are known to always exist for deterministic complete reduced FSMs but it is not the case for nondeterministic FSMs. It is also known that in this case, adaptive HS exist more often and usually are shorter than the preset. Nowadays, a number of specifications are represented by nondeterministic FSMs and thus, a deeper study of such sequences is required. There exist sufficient and necessary conditions for the existence of an adaptive HS for complete nondeterministic FSMs when each state can be an initial state but those conditions become only sufficient for weakly initialized FSMs where only some states are initial. In this paper, we propose sufficient and necessary conditions for a weakly initialized FSM to have an adaptive homing sequence, possibly up to given length, which are based on deriving an appropriate so-called homing FSM. The experimental evaluation of the existence of adaptive and preset HS is performed for randomly generated FSMs.

Keywords—Finite State Machine (FSM), weakly initialized FSM, adaptive homing sequence

I. INTRODUCTION

There is a big body of work for deriving tests with guaranteed fault coverage using Finite State Machine (FSM) based test derivation methods [see, for example, 1-4]. In order to reduce the size of a returned test suite so-called state identification sequences are utilized [5]. When the researchers minimize the number of resets in a test suite, notions of homing and synchronizing sequences (HS and SS) [6-8] are used. These sequences indicate the current state of a system under test and can be preset or adaptive [5]. Preset input sequences are derived before starting the identification procedure, while for adaptive sequences, the next input can depend on the outputs produced for the sequence of previous inputs. In order to derive a preset HS, a successor tree of an FSM under investigation is usually constructed [9, 10] and those techniques exist for deterministic and nondeterministic, partial and complete, weakly initialized and non-initialized FSMs [11]. A completely different QBF approach for deriving a preset HS is proposed in [12]. An adaptive sequence is usually represented by a tree or an acyclic FSM [10, 13, 14] called a *test case*. For nondeterministic FSMs

adaptive homing and synchronizing sequences exist more often than the preset and usually are shorter. As shown in [15], such sequences become useful in passive testing for minimizing checking efforts at each step.

There are sufficient and necessary conditions for existence and derivation of adaptive and preset homing sequences when a nondeterministic FSM is non-initialized, i.e., each state can be an initial state [10, 17, 18]. The problem of deriving an HS for weakly initialized machines is harder as it is known that checking the existence of a preset homing sequence for weakly initialized FSMs is PSPACE-complete [16, 18]. Correspondingly, researchers propose a number of heuristics to derive an HS of reasonable length [see, for example, 19]. When a nondeterministic FSM is complete and non-initialized there exists an adaptive HS if and only if such a sequence exists for each pair of states, moreover, length of such sequence is known to be polynomial with respect to the number of FSM states if it exists [17]. When an FSM is weakly initialized the above condition becomes only sufficient. The approach proposed in the same paper for checking sufficient and necessary conditions for the existence of an adaptive homing sequence for a weakly initialized complete FSM is based on enumerating homing subsets of states starting from state pairs until the set containing all initial states is obtained or no homing subsets can be further derived. Correspondingly, the approach becomes rather complicated for machines which have many states. In this paper, we suggest another formal approach for deriving an adaptive HS that is based on deriving an appropriate homing FSM that in fact, extends the method proposed in [19] and follows the approach proposed in [20] for deriving adaptive distinguishing sequences. In this paper, there are two techniques for deriving an adaptive HS. First, a homing FSM with the initial state that is the set of all initial states is derived and similar to distinguishing machines, it is shown that an adaptive HS exists if and only if the homing FSM has no complete submachine. According to our experiments with distinguishing machines, the homing FSM can be rather big [22], and thus, we also propose to construct its submachine for checking if there exists an adaptive HS up to given length. The experimental results for evaluation of the existence of adaptive and preset homing sequences for randomly generated FSMs are presented.

The rest of the paper has the following structure. Section II contains the preliminaries. Section III presents the procedures for deriving a homing FSM and an adaptive homing sequence based on the homing FSM. Section IV contains experimental results and Section V concludes the paper.

II. PRELIMINARIES

In this section, we briefly introduce definitions and notations which are mainly taken from the papers [11, 17, 22].

A. Finite State Machines

A *Finite State Machine* (FSM), or simply a *machine*, is a 5-tuple $S = (S, I, O, h_S, S_{in})$ where S is a finite non-empty set of states with the set S_{in} of initial states, I and O are finite input and output alphabets, and $h_S \subseteq S \times I \times O \times S$ is a *transition relation*. FSM S is *noninitialized* if $S_{in} = S$. If $|S_{in}| = 1$ then the FSM is an *initialized* FSM; otherwise, the FSM is *weakly initialized*. FSM S is *nondeterministic* if for some pair $(s, i) \in S \times I$, there exist several pairs $(o, s') \in O \times S$ such that $(s, i, o, s') \in h_S$; otherwise, the FSM is *deterministic*. FSM S is *complete* if for each pair $(s, i) \in S \times I$ there exists $(o, s') \in O \times S$ such that $(s, i, o, s') \in h_S$; otherwise, the FSM is *partial*. FSM S is *observable* if for every two transitions $(s, i, o, s_1), (s, i, o, s_2) \in h_S$ it holds that $s_1 = s_2$. In the following, we consider complete observable possibly nondeterministic FSMs if the contrary is not directly stated. An example of a complete nondeterministic FSM is shown in Figure 1 where states 1, 2, 3 are initial states.

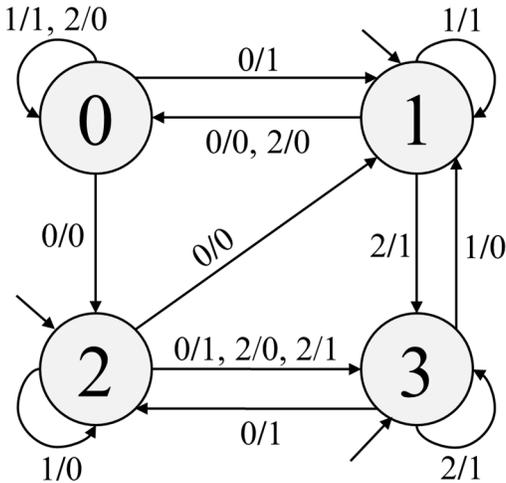


Fig. 1. Complete nondeterministic FSM S

Given $i \in I$, $o \in O$ and a state s of a complete observable FSM S , state s' is the io -successor of state s if $(s, i, o, s') \in h_S$. Such an io -successor of s not necessary exists and in this case, we say that the io -successor of state s is the empty set. A *trace* of FSM S at state s is a sequence of input/output pairs which label consecutive transitions starting from state s . Given a trace $\gamma = i_1/o_1 \dots i_l/o_l$ of FSM S , a sequence $i_1 \dots i_l$ is an *input sequence* of the trace γ , a sequence $o_1 \dots o_l$ is an *output sequence*; γ -successor of state s is a state which is reached from s via the trace γ . If γ is not a trace at state s then the γ -successor of state s does not exist. Given the set S_{in} of initial states, the γ -successor of S_{in} is the set of γ -successors over all states of the set S_{in} .

B. Homing Test Case definition

An input sequence α is *adaptive* if the next input depends on the output to the previous one. Such an input sequence can be represented by a special FSM called a *test case* [13]. Given an input alphabet I and an output alphabet O , a *test case* $TC(I, O)$ (over alphabets I and O) is an initialized observable FSM $T = (T, I, O, h_T, t_0)$ with an acyclic transition graph; each state is either a deadlock state or only one input with all possible outputs is defined at the state. Correspondingly, if $|I| > 1$ then a test case is a partial FSM. Given a complete FSM S over alphabets I and O , a test case $TC(I, O)$ represents an adaptive sequence for the FSM S and can be applied in the following way. Given a defined input i_1 at the initial state t_0 of $TC(I, O)$, this input is applied to FSM S first and $TC(I, O)$ moves to the i_1o -successor t_1 of state t_0 according to the produced output o . The procedure terminates once a deadlock state is reached. The *length* of an adaptive input sequence is the length of a longest trace from the initial state to a deadlock state of $TC(I, O)$. A test case $TC(I, O)$ is a *homing test case* (HTC) for an FSM S if for every trace γ of $TC(I, O)$ from the initial state to a deadlock state, the γ -successor of the set S_{in} is a singleton or it does not exist. Thus, the reached state can be uniquely determined based on the response of an FSM under study to adaptive homing sequences. An HTC represents an *adaptive homing sequence* and a homing test case for a machine S in Figure 1 is shown in Figure 2. The deadlock boxes have states reachable after the corresponding trace.

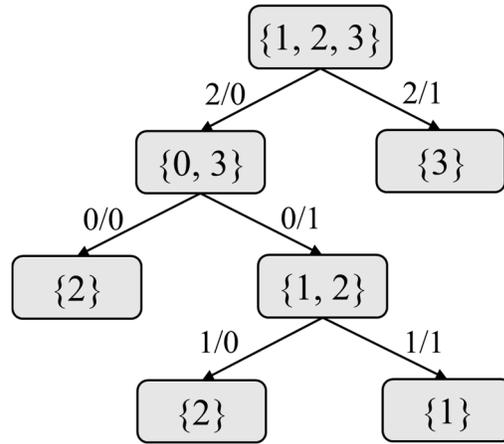


Fig. 2. HTC T for the FSM S

III. TEST CASE DERIVATION

The method for deriving an HTC based on a homing machine for an FSM with two initial states is proposed in [20]. In this paper, we extend this approach to a weakly initialized FSM with any subset of initial states. Given a complete observable nondeterministic FSM $S = (S, I, O, h_S, S_{in})$, a homing FSM S_{home} is derived in the following way. The set of inputs (outputs) S_{home} coincides with the set of inputs (outputs) of S , while states S_{home} are defined over the set of non-empty and non-singleton subsets of states of S and the homing machine can have a special state F (*FAIL*). The process of constructing transitions of S_{home} starts at the initial state which corresponds to the set S_{in} of initial states of FSM S and the machine $S_{home} = (S', I, O, h_{S(home)}, \{S_{in}\})$ is a minimal machine that can be constructed using the rules listed below.

Given a state b of S_{home} , let b be a non-empty subset of states of S that is not a singleton.

- 1) There is transition (b, i, o, b') in S_{home} if and only if b' is not a singleton and b' is the non-empty io -successor of the set b . In this case, state b' is added to the set of states of S_{home} .
- 2) There is transition (b, i, o', F) in S_{home} if and only if there exists $o' \in O$ such that the non-empty io' -successor of b is the set b . In this case, state F is added to the set of states of S_{home} .
- 3) A transition from state b under input i in S_{home} is labeled as an undefined transition if and only if each io -successor of the set b is a singleton or does not exist.
- 4) At the state F , there is a transition (F, i, o, F) for every io pair, $i \in I, o \in O$.

Similar to the theorem in [21] for adaptive distinguishing sequences, the following statement can be established.

Theorem 1. FSM S has an adaptive HS if and only if the FSM S_{home} has no complete submachine.

Sketch of the proof. In [17], it is proven that there exists an HTC for FSM if and only if the set S_{in} is k -homing for some $k > 0$. A subset b is 1-homing if there exists an input i_b such that for every $o \in O$, the $i_b o$ -successor of the set b has at most one state. Let all the k -homing subsets be determined for some $k > 0$. Then a subset c is $(k+1)$ -homing if it is k -homing and there exists an input i such that for every $o \in O$, the io -successor of c is at most k -homing or has at most one state. The subset b is *homing* if b is k -homing for some $k > 0$.

We now show that the set S_{in} is k -homing for some $k > 0$ if and only if the machine S_{home} has no complete submachine.

\Rightarrow Let the set S_{in} be k -homing for some $k > 0$. Given the machine S_{home} , if the initial state is 1-homing then there is an undefined transition at this state and thus, the S_{home} has no complete submachine. Suppose that the statement holds if the set S_{in} is m -homing for some $m > 0$.

Induction step. If the set S_{in} is $(m+1)$ -homing then by definition, there exists an input i such that for each output o , the io -successor S' of S_{in} is at most m -homing or has at most one state. Consider submachines S_1, \dots, S_r of S_{home} with the initial states $\{S_1\}, \dots, \{S_r\}$ which are io -successors of S_{in} . According to the induction assumption, each machine S_1, \dots, S_r has no complete submachine. Correspondingly, after deleting states S_1, \dots, S_r from S_{home} there will be an undefined transition under some input at the initial state, i.e., the FSM S_{home} has no complete submachine.

\Leftarrow Suppose now that there is no complete submachine in S_{home} . The latter means that after iterative deleting states with undefined transitions from S_{home} there will be an undefined transition at the initial state and correspondingly, the set S_{in} is m -homing for some $m > 0$.

The check whether the FSM S_{home} has a complete submachine can be performed by iterative deleting states with

undefined transitions from S_{home} . By definition of a homing FSM S_{home} , an input i is an *undefined* input at state b if there are no transitions from state b under i , i.e., if each non-empty io -successor of the set b is singleton. States with undefined inputs are iteratively removed from S_{home} with all incoming transitions until either there are no undefined inputs in S_{home} or the initial state of S_{home} has an undefined input. When deleting a state b with undefined inputs, we denote by $i(b)$ some undefined input. The initial state has an undefined input if and only if FSM S has an HTC T that can be derived from S_{home} by using saved undefined inputs when removing states. In this case, the initial state t_0 of T corresponds to the initial state of S_{home} that is $b_0 = S_{in}$. The input $i(b_0)$ is the only defined input at state t_0 of T and T has a transition $(t_0, i(b_0), o, t)$ if and only if FSM S_{home} has a transition $(S_{in}, i(b_0), o, t)$. If for some $o \in O$, there is no transition $(S_{in}, i(b_0), o, t)$ in S_{home} , then transition $(S_{in}, i(b_0), o, DL)$ is added to FSM T where DL is the deadlock state. In the same way, transitions for every state t of HTC T are constructed. For the FSM S in Figure 1, the fragment of corresponding homing machine S_{home} is presented in Figure 3 and an HTC of height two can be constructed by the iterative removal of states with undefined transitions (Figure 4).

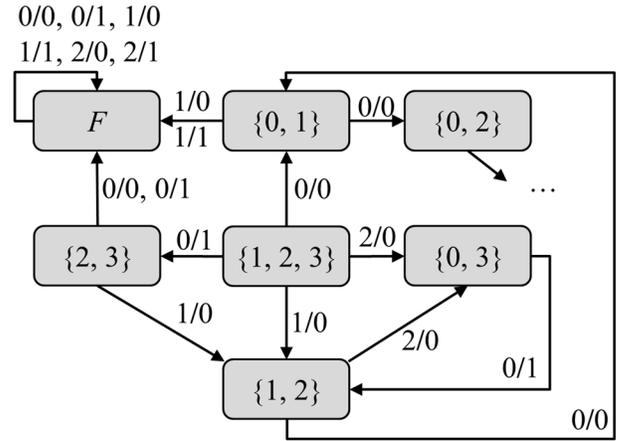


Fig. 3. Homing FSM S_{home} for the FSM S in Fig. 1

In fact, the above description of an HTC based on iterative deleting states from the FSM S_{home} does not guarantee that an adaptive homing sequence of minimal length is constructed. For deriving an adaptive homing sequence of minimal height, the following rule could be added: the construction of the homing FSM has to be limited with states which are reachable from the initial state by an input sequence of length up to $L > 0$. All transitions from states reached only by a sequence of length L are directed to state F with all possible outputs. In order to derive a shortest adaptive homing sequence, the machine S_{home}^L will be constructed starting from $L = 1$. Similar to the previous theorem, the following theorem can be proven.

Theorem 2. The FSM S has an HTC of length up to L if and only if the machine S_{home}^L has no complete submachine.

The same iterative procedure of deleting states with undefined inputs can be applied for checking the existence of an adaptive homing sequence of length up to L as well as for deriving a corresponding HTC (if it exists). The following

procedure can be proposed for deriving an HTC for a shortest adaptive homing sequence (if there exists an HTC of length up to L).

Algorithm 1. Deriving an adaptive homing sequence of minimal length

Input. A complete observable possible nondeterministic FSM S with n states, integer $L > 0$

Output. An HTC representing a shortest adaptive HS of length up to L for FSM S or the message ‘FSM S has no adaptive homing sequence of length up to L ’

$l := 1;$

The homing FSM S^1_{home} has the initial state S_{in} , state F and states which are reachable from S_{in} by a single input; all transitions from states reached from S_{in} by a single input are directed to state F with all possible outputs.

If there is an undefined input at the initial state

Then output the message ‘FSM S has a homing sequence of length one’, derive a corresponding HTC for any undefined input at state S_{in} and **END** the procedure.

Else

While $l < L$

$l++;$

Derive a homing FSM S^l_{home} : for each state $b \neq F$ of S^{l-1}_{home} which is reachable from S_{in} only by trace of length $l-1$, add to S^l_{home} new states which are reachable from state b by a single input and corresponding transitions for such inputs; all transitions from states reachable from S_{in} only by a trace of length l are directed to state F with all possible outputs.

If $S^l_{home} = S^{l-1}_{home}$

Then output the message ‘FSM S has no adaptive homing sequence of length up to L ’ and **END** the procedure.

Else

If S^l_{home} has no complete submachine

Then derive an HTC for a corresponding adaptive homing sequence iteratively deleting states with an undefined input;

Output the message ‘the obtained HTC represents a shortest HS for FSM S ’ and **END** the procedure.

Output the message ‘FSM S has no adaptive homing sequence of length up to L ’ and **END** the procedure.

Theorem 3. Given integer $L > 0$ and a complete observable possibly nondeterministic FSM S , if FSM S has an adaptive HS of length up to L , then Algorithm 1 returns an HTC for a shortest adaptive homing sequence of FSM S .

Indeed, due to Theorem 2, there exists an adaptive homing sequence of length up to l if and only if the FSM S^l_{home} has no complete submachine, i.e., at the l -th iteration the initial state of the homing machine S^l_{home} has an undefined input after iterative deleting states with undefined inputs. As we start with $l = 1$ and the procedure terminates once there is an HTC of length $l \leq L$, the procedure returns an HTC for an adaptive homing sequence of length up to L if such a sequence exists.

By direct inspection, one can assure that FSM S^1_{home} for a FSM S in Figure 1 has only defined transitions at the initial state and six states. S^1_{home} has complete submachine and thus, the FSM S has no adaptive homing sequence of length 1. FSM

S^2_{home} for a FSM S has seven states and is presented in Figure 3 if transitions from state $\{0, 2\}$ are directed to state F with all outputs (state $\{0, 2\}$ could be added only on the second iteration). S^2_{home} has no complete submachine and thus, FSM S has an adaptive homing sequence of length 2 represented by a homing test case in Figure 4. Note that S^2_{home} does not coincide with S_{home} and in general, a complete homing FSM can have significantly more states than FSMs S^l_{home} for relatively small values of L .

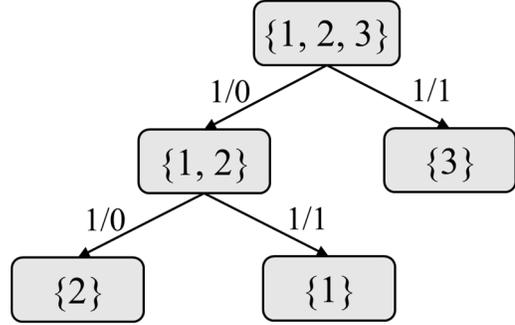


Fig. 4 An HTC for FSM S in Fig. 1

IV. EXPERIMENTAL RESULTS

For our experiments we used randomly generated complete observable nondeterministic FSMs with two inputs and two outputs. For each generated FSM, shortest preset and adaptive HSs were constructed if such sequences exist. The number of FSM transitions for each pair $(s, i) \in S \times I$ was at most two. The results of the experiments are given in Table 1. For each number of states 200 nondeterministic FSMs were randomly generated and the number of noninitialized machines which have a preset HS (Column 2) and those which have an adaptive HS (Column 3) were calculated. As it can be observed, differently from deterministic complete FSMs not every randomly generated noninitialized nondeterministic FSM has a HS; moreover, the percentage of such machines increases when the number of states increases.

TABLE 1. EVALUATION OF NUMBER OF FSM WITH PRESET AND ADAPTIVE HS

Number of states	Number of FSMs with a preset HS	Number of FSMs with an adaptive HS
4	36	52
5	106	124
6	64	87
7	193	195

For example, 193 FSMs of 200 randomly generated FSMs with 7 states have a preset HS. The number of machines which have an adaptive HS is bigger but more experiments are needed to evaluate the number of such machines depending on the number of states and/or some features of such machines.

When talking about length of an adaptive HS, we should say that adaptive sequences are not much shorter than the preset but on average, length of adaptive sequences increases when an FSM has no preset HS. Again, more experiments should be

performed in order to evaluate length of adaptive HSs for nondeterministic FSMs.

V. CONCLUSIONS

In this paper, we have proposed an approach for deriving adaptive homing sequences for complete observable possibly nondeterministic FSMs. Experiments were conducted for the evaluation how often preset and adaptive homing sequences exist for nondeterministic FSMs with two inputs and two outputs. However, the complexity of checking if an adaptive homing sequence exists for a weakly initialized FSM as well as the complexity of deriving such a sequence (if it exists) still remains unknown and needs more research.

ACKNOWLEDGMENT

This work is partly supported by RFBR project N 19-07-00327/19.

REFERENCES

- [1] T. S. Chow: Test design modeled by finite-state machines. *IEEE Transactions on Software Engineering*, 4(3): Pp. 178-187 (1978)
- [2] G. Bochmann, , and A. Petrenko: Protocol testing: review of methods and relevance for software testing. In *Proc. of International Symposium on Software Testing and Analysis*, Seattle, Pp. 109-123 (1994)
- [3] D. Lee, M. Yannakakis: Testing finite-state machines: state identification and verification. *IEEE Trans. on Computers*, 43(3): Pp. 306-320 (1994)
- [4] R. Dorofeeva, K. El-Fakih, S. Maag, A. Cavalli, N. Yevtushenko: FSM-based conformance testing methods: a survey annotated with experimental evaluation. *Information and Software Technology*, 52: Pp. 1286-1297 (2010)
- [5] Gill A. *Introduction to the Theory of Finite-State Machines*, 1964, 272 p.
- [6] F.C. Hennie. Fault-detecting experiments for sequential circuits. In: *Proceedings of Fifth Annual Symposium on Circuit Theory and Logical Design*, Pp. 95–110 (1965)
- [7] Jourdan, G.-V., Ural, H., and Yenigun, H., "Reduced checking sequences using unreliable reset", *Information Processing Letters*, Vol.115, No.5, Pp.532-535 (2015)
- [8] H. Ural, F. Zhang, and J.C. Zhang, "Effects of overlapping subsequences in constructing checking sequences", *Journal of Advances in Information Sciences*, Vol.1, No.1, Pp. 59-73 (2013)
- [9] Kohavi, Z.: *Switching and Finite Automata Theory*. McGraw-Hill, New York (1978)
- [10] N. Kushik. *Methods for deriving homing and distinguishing experiments for nondeterministic FSMs*. PhD thesis, Tomsk State University, 137 p.(2013)
- [11] H. Yenigun, N. Yevtushenko, N. Kushik, J. López: The effect of partiality and adaptivity on the complexity of FSM state identification problems. *Trudy ISP RAN/Proc. ISP RAS*, 30 (1), Pp. 7-24 (2018)
- [12] Hung-En Wang, Kuan-Hua Tu, Jie-Hong R. Jiang, Natalia Kushik: *Homing Sequence Derivation with Quantified Boolean Satisfiability*. – *Lecture Notes in Computer Science (LNCS)*, № 10533, Pp. 230-242 (2017)
- [13] A. Petrenko, N. Yevtushenko: *Adaptive Testing of Deterministic Implementations Specified by Nondeterministic FSMs*. *Lecture Notes in Computer Science (LNCS)*, № 7019, Pp. 162-178 (2011)
- [14] Hüsnu Yenigün, Nina Yevtushenko, Natalia Kushik: Some classes of finite state machines with polynomial length of distinguishing test cases. In *Proceedings of SAC 2016*, Pp. 1680-1685 (2016)
- [15] N. Kushik, J. López, A. Cavalli, N. Yevtushenko: Improving Protocol Passive Testing through "Gedanken" Experiments with Finite State Machines. In *Proceedings of QRS 2016*. Pp. 315-322 (2016)
- [16] S. Sandberg: *Homing and Synchronization Sequences*. *Model Based Testing of Reactive Systems*, LNCS № 3472, Pp. 5-33 (2005)
- [17] N. Kushik, K. El-Fakih, N. Yevtushenko: *Adaptive Homing and Distinguishing Experiments for Nondeterministic Finite State Machines*. *Lecture Notes in Computer Science (LNCS)*, № 8254, Pp. 33-48 (2013)
- [18] N. Kushik, V. Kulyamin, N. Evtushenko: On the complexity of existence of homing sequences for nondeterministic finite state machines. *Programming and Computer Software*, 40(6): Pp. 333-336 (2014)
- [19] N. Kushik, H. Yenigün: *Heuristics for Deriving Adaptive Homing and Distinguishing Sequences for Nondeterministic Finite State Machines*. *Lecture Notes in Computer Science (LNCS)*, № 9447, Pp. 243-248 (2015)
- [20] N. Kushik, N. Yevtushenko: Adaptive Homing is in P. – *Electronic Proceedings in Theoretical Computer Science*, 180, Pp. 73-78 (2015)
- [21] El-Fakih, K., Yevtushenko, N., Kushik, N.: Adaptive distinguishing test cases of nondeterministic finite state machines: Test case derivation and length estimation. *Formal Aspects of Computing*, 30(2): Pp. 319-332 (2018)
- [22] A. Tvardovskii, N. Yevtushenko: *Deriving adaptive distinguishing sequences for Finite State Machines*. *Trudy ISP RAN/Proc. ISP RAS*, 30 (4), Pp. 139-154 (2018)

Non-Canonical Topography of the z-Plane Discretized due to Quantization of the IIR Digital Filter Coefficients

Vladislav Lesnikov¹, Tatiana Naumovich², Alexander Chastikov³
Department of radioelectronic systems,
Vyatka State University

Kirov, Russia

¹vladislav.lesnikov.ru@ieee.org, ²ntv_new@mail.ru, ³alchast@mail.ru

Abstract— It is known that all possible positions of the zeros and poles of IIR digital filters with finite word length form some discrete structure in z-plane. In this paper, this structure is called the z-plane topography. Discretization of the z-plane is explained by the fact that, due to quantization of the filter coefficients, not any of its points can be a zero or a pole. In their previous publications, the authors showed that for practically implemented digital filters, zeros and poles are elements of the algebraic number set. It was also shown that the topography of the sampled z-plane is completely determined by the degree of algebraic numbers and the bitness of fractional part of the transfer function coefficients. The corresponding topography is proposed to be called canonical. At the same time, structures are known for which the topography differs from the canonical one. For them, the name of non-canonical topography is suggested. These structures include the structure known as the coupled form, the normal form, the structure of Gold and Raider. This article is devoted to the study of such structures. It may seem that this is a completely new topography, not connected with the canonical one. However, this is not true. This paper establishes the relationship between the canonical topography and the topography of the coupled and other structures. It is shown that the coefficients of such structures are not independent. They are related by some additional equations. As a result, the number of possible zeros and poles is reduced. But the remaining zeros and poles are elements of the set that make up the canonical topography.

Keywords— IIR digital filters, Possible pole-zero locations, Grid of allowable pole and zero positions, Variation curve of the poles and zeros, Quantization of pole locations, Plane algebraic curves

I. INTRODUCTION

Despite the fact that the theory of designing IIR digital filters has been developed for a long time, the difficulties accompanying the practical development process in the case of stringent requirements for characteristics turn out to be so complex that developers have to abandon IIR filters in favor of FIR filters. Overcoming the problems arising in this case requires a deep study of the nature of IIR filters with a finite word length (FWL).

The traditional approach to the synthesis of IIR digital filters includes the stages of functional and structural synthesis.

The work was supported by a grant from the Russian Foundation for Basic Research 18-07-00986.

Functional synthesis involves the calculation of the zeros and poles of the transfer function without taking into account the FWL. The finite bit depth of the filter coefficients is taken into account only at the stage of structural synthesis. Therefore, the results of structural synthesis distort the results of functional synthesis and, therefore, errors are introduced into the filter characteristics. Accounting for these circumstances complicates the procedures necessary to meet the requirements for filter characteristics.

The new paradigm of the FWL digital filter design developed by the authors, is based on the fact that the final placement of zeros and poles in the z-plane is completed even before structural synthesis (at the stage of functional synthesis) [1], [2]. In our papers [3], [4], we established that for a FWL coefficients of IIR digital filter with any structure, the zeros and poles are elements of the set of algebraic numbers. The set of algebraic numbers is the set of all possible roots of polynomials with coefficients belonging to the set of rational numbers [5]. The finite word length of the coefficients of the practicable digital filters causes the coefficients of their transfer functions to be rational numbers. Therefore, the zeros and poles of such filters are algebraic numbers. It follows that not every point of the z-plane can be the root of the polynomials of the numerator and denominator of the transfer function.

The discrete structure of the allowed positions of the poles of second-order digital filters with a given coefficient bitness has long been known [6] – [9]. The topography of a discretized z-plane was studied in detail in our paper [10] for poles of second-order filters. An attempt is made in [11] – [13] to investigate higher order filters.

All results obtained in [10]-[13] are characterized by the fact that the coefficients of a particular structure can take on any value that is permissible at a given capacity. The corresponding topography will be called canonical.

In this case, the topography is determined only by the degree of the corresponding algebraic numbers and the digit capacity of the fractional part of the transfer function coefficients.

At the same time, structures are known for which the topography differs from the canonical one. For example, for a

coupled second order filter structure, the poles are located at the nodes of a square lattice. It may seem that this is a completely new topography, not connected with the canonical one. However, this is not true. This article establishes the relationship between the canonical topography and the topography of the coupled and other structures.

Below we will call these new topographies non-canonical. For the second-order filters with transfer function

$$H(z) = \frac{\sum_{i=0}^n b_i z^{n-i}}{z^n + \sum_{i=1}^n a_i z^{n-i}}, \quad (1)$$

it will be shown that the set of all possible poles constituting the non-canonical topography is a subset of the set of poles forming the canonical topography. Filters with non-canonical topography are characterized by the fact that some of their coefficients are related by equations, which reduces the number of degrees of freedom and reduces the power of the set of possible poles.

II. SECOND ORDER DIGITAL FILTERS WITH NON-CANONICAL POLE TOPOGRAPHY

A. Coupled Form (Normal Form, Gold and Rader Structure)

The known coupled structure [8] – [11] is shown in Fig. 1.

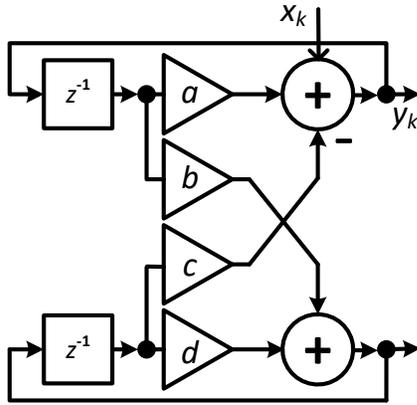


Fig. 1. Coupled form.

The transfer function of the digital filter, implemented by the coupled structure, is equal to

$$H(z) = \frac{1 - dz^{-1}}{1 - (a+d)z^{-1} + (ad+bc)z^{-2}}. \quad (2)$$

If the coefficients a , b , c , and d were independent, then the topography of the poles would be canonical and is shown in Fig. 2 small blue dots.

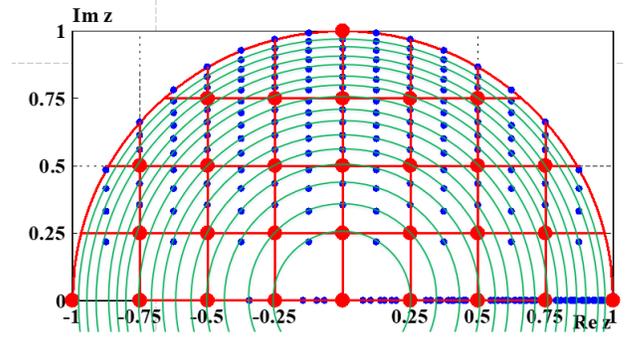


Fig. 2. Possible pole position for transfer function (2) ($m_{a1}=2$, $m_{a2}=4$).

A feature of this topography is that if the bitness of these coefficient fractional parts are the same and are equal to m , then the digit capacity of the fractional part of the coefficient

$$a_1 = -(a+d) \quad (3)$$

is also equal to $m_{a1}=m$, and the bitness of the fractional part of the coefficient

$$a_2 = ad+bc \quad (4)$$

is equal to $m_{a2}=2m$.

However, the coefficients of the coupled form are related by

$$\begin{cases} a = d, \\ b = c. \end{cases} \quad (5)$$

The transfer function becomes equal to

$$H(z) = \frac{1 - az^{-1}}{1 - 2az^{-1} + (a^2 + b^2)z^{-2}}. \quad (6)$$

The dependence between the coefficients of the structure (5) leads to the fact that there is a relationship between the coefficients a_1 and a_2 :

$$a_2 = 0.25a_1^2 + b^2 \quad (7)$$

Therefore, from the set of possible values of the coefficients a_1 and a_2 fell coefficients that do not satisfy the equation (7) (Fig. 3) and, accordingly, the number of permissible poles has decreased.

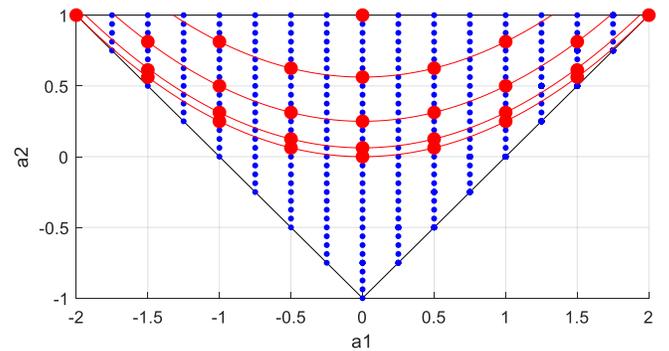


Fig. 3. The coefficients of the polynomial of denominator of the coupled form transfer function ($m_{a1}=2$, $m_{a2}=4$).

B. Other Coupled Structures

In addition to the coupled structure (Fig. 1), other structures are also known [16] that have the non-canonical topography of the z -plane (Fig. 4, Fig. 7, and Fig. 10).

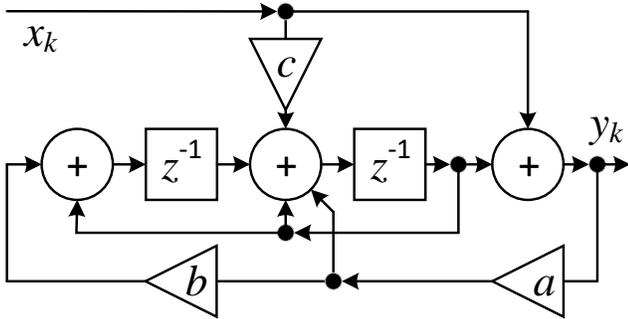


Fig. 4. Digital filter with transfer function (8).

Digital filter with the structure of Fig. 4 has a transfer function

$$H(z) = \frac{z^2 - (1-c)z - 1}{z^2 - (1+a)z - (1+ab)}, \quad (8)$$

The coefficients of the denominator of the transfer function are calculated as

$$\begin{cases} a_1 = -(1+a), \\ a_2 = -(1+ab). \end{cases} \quad (9)$$

From (9) it follows that the coefficients a_1 and a_2 are not independent:

$$a_2 = ba_1 + (b-1). \quad (10)$$

This means that the values of a_1 and a_2 , unsatisfying (10) are excluded from the set of allowed pairs (a_1, a_2) (Fig. 5).

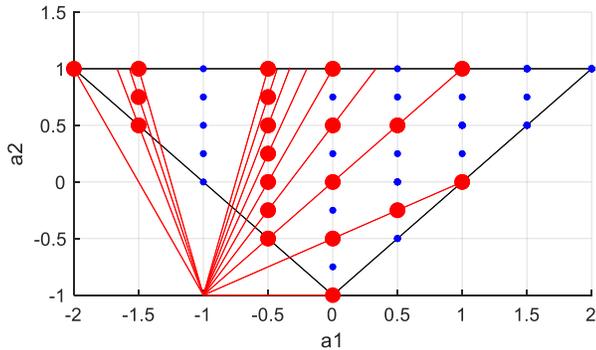


Fig. 5. The coefficients of the polynomial of denominator of the transfer function (8) ($m_{a1}=1, m_{a2}=2$).

In this case, a smaller number of poles are selected from the canonical topography to the non-canonical one (Fig. 6).

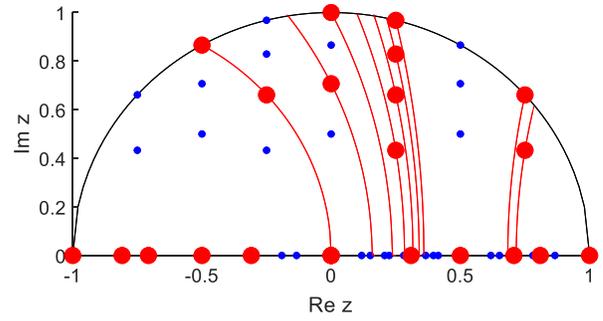


Fig. 6. Possible pole position for transfer function (8) ($m_{a1}=1, m_{a2}=2$).

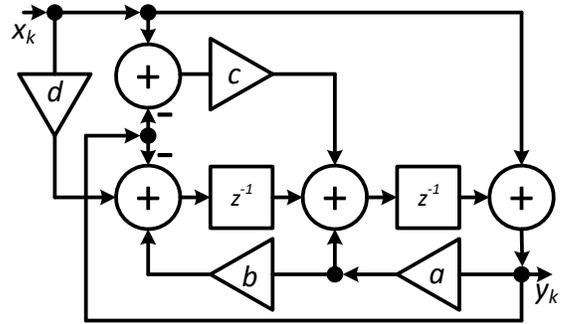


Fig. 7. Digital filter with transfer function (11).

Digital filter with the structure of Fig. 7 has a transfer function

$$H(z) = \frac{z^2 + cz + d}{z^2 - (a-c)z + (1-ab)}. \quad (11)$$

The coefficients of the denominator of the transfer function are calculated as

$$\begin{cases} a_1 = c - a, \\ a_2 = 1 - ab. \end{cases} \quad (12)$$

From (12) it follows that the coefficients a_1 and a_2 are not independent:

$$a_2 = ba_1 + (1-bc). \quad (13)$$

This means that the values of a_1 and a_2 , unsatisfying (13) are excluded from the set of allowed pairs (a_1, a_2) (Fig. 8).

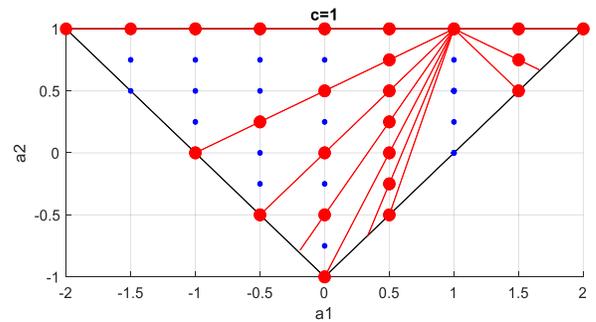


Fig. 8. The coefficients of the polynomial of denominator of the transfer function (11) ($m_{a1}=1, m_{a2}=2$).

In this case, a smaller number of poles are selected from the canonical topography to the non-canonical one (Fig. 9).

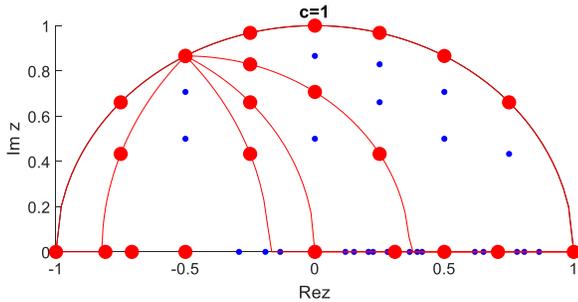


Fig. 9. Possible pole position for transfer function (11) ($m_{a1}=1, m_{a1}=2$).

III. CONCLUSIONS

Thus, the differences in the z-plane topography from the canonical one are explained by the fact that the coefficients of the digital filters are related to additional restrictions, as a result of which the power of the set of possible values of the zeros and poles of the digital filters decreases. For this reason, it is reasonable to call the structures of the filters with the noncanonical topography of the z-plane also coupled.

It is obvious that IIR digital filters for which inequality

$$N_{mpy} < N_{mpy\ can}, \quad (14)$$

where N_{mpy} is the number of multiplication blocks in the structure under consideration, $N_{mpy\ can}$ is the number of blocks of multiplication in canonical structure, is valid will also have non-canonical z-plane topography. This is explained by the fact that some values of the coefficients of the transfer function in this structure are not feasible.

REFERENCES

- [1] V. Lesnikov, A. Chastikov, T. Naumovich, and S. Armishev, "A new paradigm in design of IIR digital filters," *8th IEEE East-West Design and Test Symposium (EWDTS 2010)*, St. Petersburg, Russia, 17-20 Sept. 2010, pp. 282-285.
- [2] V. Lesnikov, A. Chastikov, T. Naumovich, and S. Armishev, "Implementation of a new paradigm in design of IIR digital filters," *8th IEEE East-West Design and Test Symposium (EWDTS 2010)*, St. Petersburg, 17-20 Sept. 2010, pp. 156-159.
- [3] V. Lesnikov, and T. Naumovich, "Number-theoretic and algebraic aspects of structural synthesis of digital filters," *Global Signal Processing (GSP 2004). The International Embedded Solutions Event*

(*The Embedded Signal Processing Conference*), Santa Clara, Ca, USA, 2004. P. 27-30.

- [4] V. Lesnikov, T. Naumovich, and A. Chastikov, "Number-theoretical analysis of the structures of classical IIR digital filters," *7th Mediterranean Conference on Embedded Computing (MECO 2018)*, Budva, Montenegro, 10-14 June 2018, 4 p., <https://doi.org/10.1109/MECO.2018.8406099>.
- [5] D. Hilbert, *The Theory of Algebraic Number Fields*, Berlin – Heidelberg – New York: Springer – Verlag, 1998.
- [6] C. J. Weinstein, Quantization Effects in Digital Filters, Technical Report 468, Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts, 21 November 1969, available: <https://www.semanticscholar.org/paper/Quantization-Effects-in-Digital-Filters-Weinstein/0e52d6dbb14fb6c137527f7919e0bc380bd276f8>.
- [7] W. Hess, *Digitale Filter: eine Einführung*, Springer Fachmedien Wiesbaden GmbH, 1993.
- [8] B. W. Bomar, "Finite Wordlength Effects," in *Digital Signal Processing Handbook*, ed. V. K. Madisetti, and D. B. Williams, Boca Raton: CRC Press LLC, 1999.
- [9] D. L. Jones, "Digital Filter Structures and Quantization Error Analysis," *Connexions*, Rice University, Houston, Texas, 2005, available: <https://cnx.org/exports/7de2003c-4c17-440b-ba29-c681e8fc14e4@1.1.pdf/digital-filter-structures-and-quantization-error-analysis-1.1.pdf>.
- [10] W. L. Mills, C. T. Mullis, and R. A. Roberts, "Normal realizations of IIR digital filters," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1979)*, Washington, DC, USA, 2-4 April 1979, Vol. 4, pp. 340 – 343, DOI: 10.1109/ICASSP.1979.1170710.
- [11] C. M. Gold, and B. Rader, *Digital Processing of Signals*, New York: McGraw-Hill, 1969.
- [12] V. Lesnikov, T. Naumovich, and A. Chastikov, "The sampling of the z-plane due to the quantization of the digital filter coefficients," *7th Mediterranean Conference on Embedded Computing (MECO 2018)*, Budva, Montenegro, 10-14 June 2018, 4 p., <https://doi.org/10.1109/MECO.2018.8405962>.
- [13] V. Lesnikov, T. Naumovich, and A. Chastikov, "Topography of z-plane which is discretized due to quantization of coefficients of digital biquad filters," *12th International Siberian Conference on Control and Communications (SIBCON 2016)*, Moscow, Russia, 12-14 May 2016, 4 p., <https://doi.org/10.1109/SIBCON.2016.7491812>.
- [14] V. Lesnikov, T. Naumovich, and A. Chastikov, "The topography of a third order IIR digital filter zeros and poles in the z-plane discretized due to the quantization of the direct form coefficients," *7th Mediterranean Conference on Embedded Computing (MECO 2019)*, Budva, Montenegro, 10-14 June 2019, pp. 374-377.
- [15] V. Lesnikov, T. Naumovich, A. Chastikov, and A. Metelyov, "Topography of the z-plane discretized by quantizing the coefficients of the canonical form of recursive digital filter," in *Computer Vision in Advanced Control Systems - 6*, M. Favorskaya, and L.C. Jain, Eds, in press (will be published by Springer in 2020).
- [16] E. Avenhaus, "A proposal to find suitable canonical structures for the implementation of digital filters with small coefficient word length," *Nachrichtentechn. Z.*, vol. 25, pp. 377-382, August 1972.

Modification of U-Net neural network in the task of multichannel satellite images segmentation

Vladimir Khryashchev
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
v.khryashchev@uniyar.ac.ru

Anna Ostrovskaya
People's Friendship University of Russia
Moscow, Russia
ostrovskaya_aa@rudn.university

Roman Larionov
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
r.larionov@uniyar.ac.ru

Alexander Semenov
People's Friendship University of Russia
Moscow, Russia
semenov.venture@mail.ru

Abstract — Results of training of convolutional neural network for satellite four-channel image segmentation are performed. Input images contain blue, green, red and near-infrared channels. The algorithm was trained to detect buildings and other urban areas. Modification of the U-Net neural network with two encoders was used. The values of Sorensen coefficient and Jaccard index were calculated for 16 different urban regions.

Keywords — image segmentation, satellite images, convolutional neural network, deep learning

I. INTRODUCTION

A satellite is a mechanical object that periodically rotates on the Earth orbit to performs certain functions and tasks for global media, military intelligence, satellite communications, meteorological observations, etc. At present, there are many satellites which are capable of making high-resolution images of the Earth surface. Table 1 presents some popular satellites.

TABLE I. SOME POPULAR SATELLITES

Model of satellite	Launch year	Price of images
Landsat 8	2013	Free
Sentinel	2014	Free
MODIS	2002	Free
WorldView	2016	Paid
QuickBird	2001	Paid
GeoEye-1	2008	Paid
IKONOS	1999	Paid
Jilin-1	2015	Paid
SPOT-6	2012	Paid
Gaofen-2	2014	Paid
TripleSat	2015	Paid

The features of obtaining satellite images of the Earth include the following peculiarities:

- Satellites simultaneously take photos of the same territory. The first image is black and white, the most detailed. The second one is color, with a resolution below average. It is impossible to take high-resolution

color photos, because the light is refracted in the Earth's atmosphere. In addition, it is needed to stretch the color snapshot, because in digital form it turns out less black and white image. Finally both images are combined.

- Space cameras identify colors differently, so the original satellite images do not look like natural photos due to diffraction and scattering in the Earth's atmosphere. In order to make colors normal for human perception, color correction is essential to be performed.
- Shooting conditions and camera type cause a shifting effect. In this case it is needed to implement eliminating distortions and transforming the original image into its orthogonal projection.

Satellites receive hundreds terabytes of photos every day. Therefore, the development of methods for their automatic processing is very relevant. Modern satellites are capable to make photos with spatial resolution of 3 m/pixel and less. It's possible to detect such small objects as buildings, landfills, etc. The development of technologies allowed to use methods of deep learning for satellite image segmentation [1].

This paper presents developed image processing method based on convolutional neural network (CNN). Such networks are capable for real-time detecting and classifying objects. CNN have millions of parameters that are automatically matched through the training. CNN has shown its effectiveness in different computer vision tasks [2].

Automatic segmentation is an important part of aerial image pre-processing. Today, convolutional neural networks are widely used for this task. In particular, U-Net neural network architecture has shown its effectiveness in medical image segmentation [3]. Also U-Net has shown good results in satellite image segmentation [4]. The main advantage of this architecture is that algorithm can show good results even with a small training datasets. The mathematical structure of CNN is parallel, so graphic processors units (GPU) are ideal instrument to work with CNN [5].

There are some specific requirements for satellite images segmentation, containing different urban areas[6]:

- Size and type of buildings and urban structures may significantly vary from cottages to huge city buildings. The algorithm should detect objects of any



Fig. 1. Sample images from Spacenet database

size very well. The usage of multiple encoders in neural network structure could solve this problem.

- The separation of objects with a high density of location. Algorithm should be penalized for bad separation of objects during training to improve output mask quality. This is achieved by careful selection of the loss function.
- Trained model should be invariant to rotations. This problem can be solved by data augmentation.
- Aerial images have different spatial resolution. It's important to create algorithm which has an ability to generalize perfectly.
- Algorithms should be noise robust. Images are shot in different weather situations. It is possible that there is a noise in some photos, for example haze and glare from reflective surfaces.

This article presents the results of training of developed neural network for satellite multispectral image segmentation in order to detect buildings and urban areas. Modification of the U-Net architecture with the second encoder is proposed. The research continues previous investigations [7,8].

The rest of the paper organized as follows. The second part describes image datasets. Architecture of neural network is shown in the third part. The fourth part presents results of

numerical experiments with big aerial image database containing different urban areas. Current research is summarized in the conclusion part.

II. SATELLITE IMAGES DATABASE

Neural network model was pre-trained on images of Spacenet database. WorldView-2 and WorldView-3 satellites were used to make eleven-bit photos, all images are eight-channel. The database contains subsets with marked buildings for training deep learning algorithms [9]. We used subset of Khartoum (region of Spacenet database) to pre-train developed CNN. Examples of images from Spacenet database are shown at Fig. 1.

Our model was trained on images of 16 different regions of Russian Federation. Every image has an appropriate mask marked by experts. The dataset covers about 30 square kilometres. The images of this database contain blue, green, red and near infrared (NIR) channels with a spatial resolution of 3 metres per pixel.

The developed CNN requires input images of 256×256 pixels. Because of this the launch of CNN training on each image and appropriate mask of dataset have been cut on two non-intersecting stripes. Each stripe was divided on patches of 256×256 pixels with step of 128 pixels, so paths intersected by half.

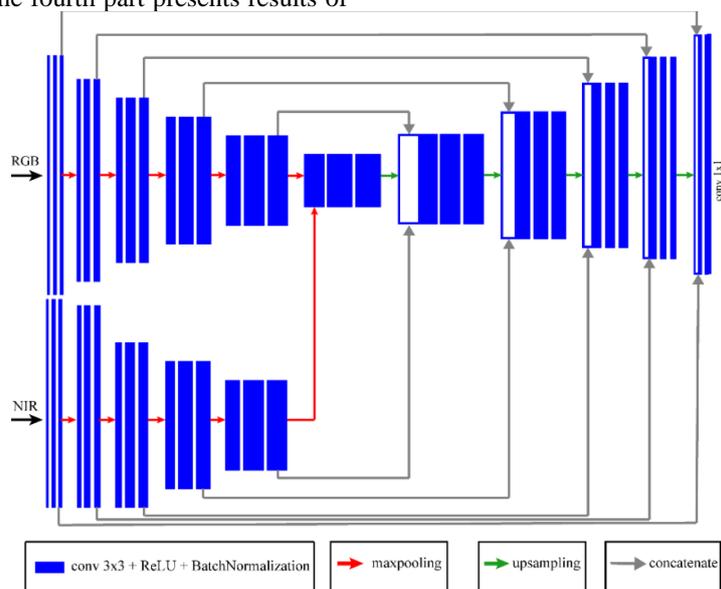


Fig. 2. U-Net neural network architecture with 2 encoders

To increase the training set, there were applied three types of data augmentation:

- Rotations on 90, 180, 270 degrees and reflections. The training set has increased 8 times after this procedure.
- Chromatic distortions. Images were translated from RGB color space to HSV color space. Random values were added to HSV coordinates of images. For the NIR channel, instead of chromatic distortions, random values from [-0.06, +0.06] interval were added. (NIR values were normalized in [0, 1] interval).
- Random shifts, scales and small degree rotations of patches.

As a result, the training image dataset contains 9784 batches each of them consist 16 images of 256×256 pixels.

III. NEURAL NETWORK ARCHITECTURE

In the current investigation the modification of well known U-Net architecture was used. Its classical structure is described in [3].

Original U-Net structure was modified. We have used two different encoders for RGB and NIR channels (fig. 2).

Developed CNN was launched on NVIDIA DGX-1 supercomputer. Adam optimizer with learning rate 0.001 was used in the learning procedure. Lovász-Softmax loss was chosen as the loss function. The training finishes after completing 100 epochs.

The comparison of training results for original U-Net with four-channel images as inputs and modification with two encoders for RGB and NIR channels is shown at fig. 3.

The outputs of encoders were concatenated before being linked with appropriate layers. As a result, the neural network has 38 convolutional layers, 37 ReLU activation functions, 37 operations of batch normalization, 1 sigmoid activation function, 10 maxpooling operations, 5 upsampling operations, 11 merge operations.

IV. NUMERICAL RESULTS

The quality of the segmentation algorithms was evaluated by Sorensen-Dice coefficient (dice) and Jaccard index (IoU).

The coefficients can be calculated by following formulas (1-4), where x and y are values of pixels, X is marked by experts mask, Y is algorithm prediction:

$$dice = \frac{2I}{S}, \quad (1)$$

$$I = \sum_{x \in X} \sum_{y \in Y} xy, \quad (2)$$

$$S = \sum_{x \in X} \sum_{y \in Y} (x + y), \quad (3)$$

$$IoU = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}, \quad (4)$$

During preliminary training on the Spacenet dataset, the value of the Sorensen coefficient reached a value of 0.84, a Jaccard index of 0.77. The values of the Sorensen coefficient and the Jaccard index for all 16 regions for the algorithm with pre-training and without pre-training are shown at Fig. 4.

Example of input image and the result of segmentation is shown on Fig. 5. Sorensen coefficient and Jaccard index are equal 0.78 and 0.67 respectively.

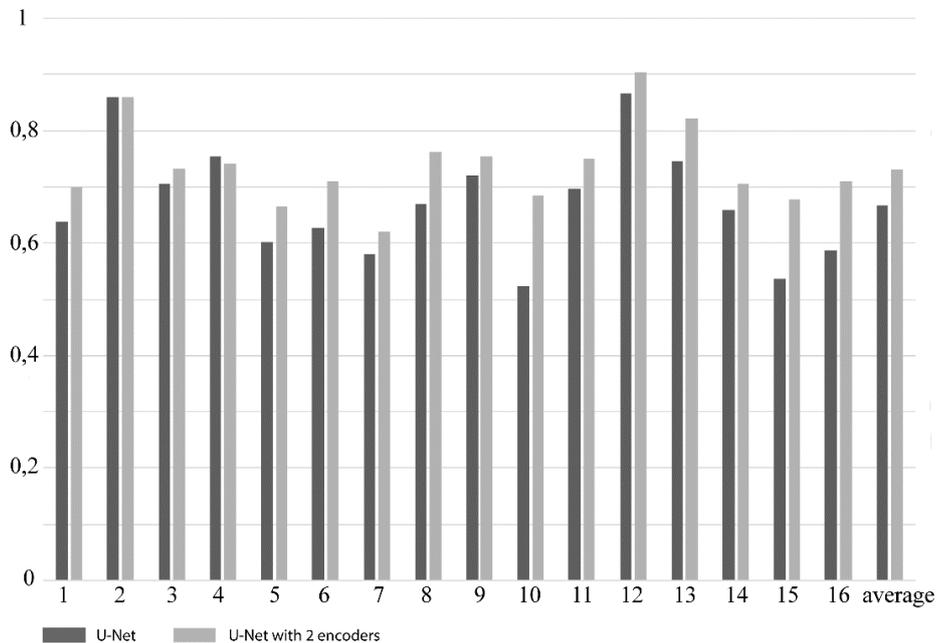


Fig. 3. Values of Sorensen coefficients for original U-Net and modified U-Net for images of 16 regions

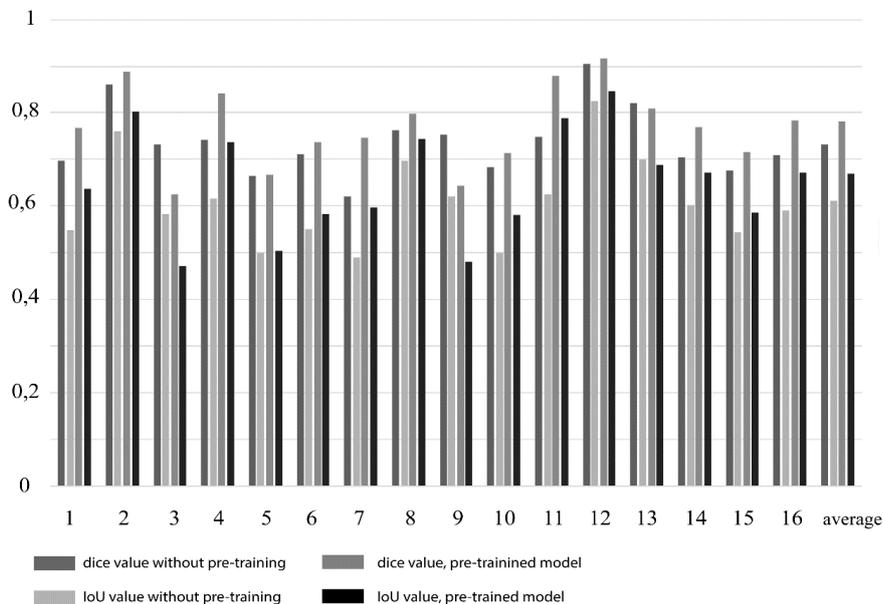


Fig. 4. Values of Sorensen coefficient and Jaccard index of pre-trained model and the algorithm without pre-training



Fig. 5. Example of input image and result of segmentation for developed algorithm

V. CONCLUSIONS

This paper describes the training process of developed convolutional neural network intended for the segmentation of satellite images. U-Net architecture with two encoders was proposed to work with four-channel images. The algorithm was pre-trained on Spacenet image dataset. The value of Sorensen coefficient is equal 0.78, Jaccard index is 0.67.

Future directions of research may include:

- Categorization of buildings based on area size and encoders for separate classes.
- Usage of object boundaries as the third class for detection.

Developed algorithm can be used for assessing the level of urbanization of various regions and tracking the construction of large objects.

ACKNOWLEDGMENT

This article was prepared with the financial support of the Ministry of science and education of the Russian Federation under the agreement No. 075-15-2019-249 from 04.06.2019 (identifier works RFMEFI57517X0167).

REFERENCES

- [1] Zhang, Liangpei, Lefei Zhang, and Bo Du. "Deep learning for remote sensing data: A technical tutorial on the state of the art." *IEEE Geoscience and Remote Sensing Magazine* (2016): 22-40.
- [2] S. Seferbekov, V. Iglovikov, A. Buslaev, A. Shvets. Feature Pyramid Network for Multi-Class Land Segmentation. Web: <https://arxiv.org/pdf/1806.03510.pdf>.
- [3] O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, LNCS, vol. 9351, 2015, pp. 234–241.
- [4] Zhang, Z., Liu, Q., & Wang, Y. (2018). "Road extraction by deep residual u-net". *IEEE Geoscience and Remote Sensing Letters*, 15(5), pp. 749-753.
- [5] A. Gulli, S. Pal., *Deep Learning with Keras*, Packt Publishing, 2017, 320 p.
- [6] X. X. Zhu et al., "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," in *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8-36, Dec. 2017.
- [7] V. V. Khryashchev, V. A. Pavlov, A. Priorov and A. A. Ostrovskaya, "Deep Learning for Region Detection in High-Resolution Aerial Images," 2018 IEEE East-West Design & Test Symposium (EWDTS), Kazan, 2018, pp. 1-5.
- [8] L. Ivanovsky, V. Khryashchev, V. Pavlov and A. Ostrovskaya, "Building Detection on Aerial Images Using U-NET Neural Networks," 2019 24th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 2019, pp. 116-122.
- [9] SpaceNet Database, Web: <http://explore.digitalglobe.com/spacenet>

Complementary JFETs Integrated into the Microwave Complementary Bipolar Double Self-Aligned Technology

Dmitry G. Drozdov
JSC "S&PE" "PULSAR", RTU MIREA
Moscow, Russia
drozdov_dmitrii@pulsarnpp.ru

Nikolay N. Prokopenko
Department "Information Systems
and Radio Engineering"
Don State Technical University
Rostov-on-Don, Russia
prokopenko@sssu.ru

Evgeny M. Savchenko
JSC "GZ "Pulsar", JSC "S&PE
"PULSAR", RTU MIREA
Moscow, Russia
savchenko@pulsarnpp.ru

Andrey I. Grushin
JSC "S&PE" "PULSAR"
Moscow, Russia
grushin_ai@pulsarnpp.ru

Pavel A. Dukanov
JSC "S&PE" "PULSAR"
Moscow, Russia
dukanov@pulsarnpp.ru

Abstract—The article presents the results of the development of complementary junction field-effect transistors, integrated into the microwave complementary bipolar technological process. The formation modes of the p-channel region of the field-effect transistor, capable of providing a low value of the cutoff voltage and high values of the shorted-gate drain current, are determined. The necessity of applying the methods of suppressing the back diffusion of boron to ensure the cutoff voltage symmetry of the junction field-effect transistors with p- and n-channels is shown.

Keywords— *Complementary Junction Field-Effect Transistors; Microwave Complementary Bipolar Transistors; Technological Process*

I. INTRODUCTION

Analog integrated circuits, based on the complementary bipolar transistors together with junction field-effect transistors (JFET), have the following advantages [1]:

- high input resistance;
- low input currents;
- low noise current, etc.

The inclusion of the complementary junction field-effect transistors into the microwave complementary bipolar technological process allows developing ICs suitable for operation at low temperatures and under the influence of ionizing radiation [2-6]. However, with such integration, a number of structural and technological difficulties arise, primarily due to the providing symmetrical values of the parameters of complementary transistors. The companies like Texas Instruments, STMicroelectronics, etc. are involved in creation of technological processes with complementary junction field-effect transistors [7-10]. Their processes are characterized by a cutoff voltage of about 1 V and a shorted-gate drain current of about 1.5 $\mu\text{A} / \mu\text{m}$.

In this paper, we will present the research findings on the integration of complementary junction field-effect transistors with symmetric static parameters into the microwave complementary bipolar double self-aligned process [11].

II. MICROWAVE COMPLEMENTARY BIPOLAR TECHNOLOGICAL PROCESS

The main features of the microwave complementary bipolar process are [12]:

- deep trench isolation;
- double implantation to form the gradient impurity profile in the collector of the pnp-transistor;
- sequential doping of layers for individual areas of complementary bipolar transistors;
- LOCOS-isolation of active regions of transistors;
- polysilicon contacts to the areas of the base and emitter;
- self-alignment of the base and emitter areas;
- silicon nitride spacers.

The use of double self-alignment and deep trench isolation enables to reduce the size of transistors and improve their dynamic characteristics at process feature size of 1 micron. The main parameters of the complementary bipolar transistors are shown in Table 1.

TABLE I. MAIN PARAMETERS OF MICROWAVE COMPLIMENTARY BIPOLAR TRANSISTORS

Parameter	Measurement mode	npn	pnp
Gain	$U_{CE} = 2 \text{ V}$	181	61
Cutoff frequency, GHz	$U_{CE} = 2 \text{ V}$	12.6	10.7
Maximum frequency of oscillation, GHz	$U_{CE} = 3 \text{ V}$	31	33
Breakdown voltage of base-collector, V	$I_C = 1 \mu\text{A}$	41	27
Breakdown voltage of collector-emitter, V	$I_C = 10 \mu\text{A}$	15.3	14.7

III. P-CHANNEL JUNCTION FIELD-EFFECT TRANSISTOR

Fig. 1 shows a cross-sectional view of an integrated p-channel junction field-effect transistor. To ensure the breakdown voltage of the drain-source of more than 15 V, a

The study has been carried out at the expense of the grant from the Russian Science Foundation (Project No. 16-19-00122-P).

design was used in which the channel region was formed by a collector of the pnp-transistor [13,14].

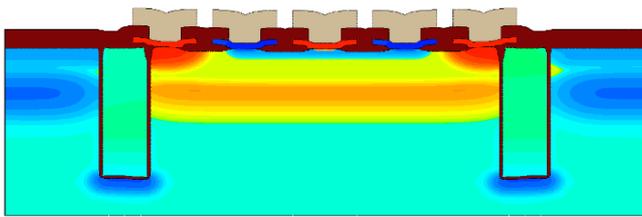


Fig. 1. The design of the integrated p-channel JFET.

The upper gate is formed on the basis of the polysilicon region of the n-type passive base and has a length $L_G = 1 \mu\text{m}$, determined by the minimum distance between the LOCOS-isolation regions.

When using a collector of the pnp-transistor, formed by double implantation, high cutoff voltage $U_{GS(off)} > 7 \text{ V}$ is observed. To reduce the cutoff voltage, it is necessary to reduce the channel thickness, which is possible only when using the top implantation of the collector of the pnp-transistor. The dependence of parameters of the transistor (cutoff voltage, shorted-gate drain current) on the implantation modes of the upper p- region is shown in Fig. 2 and Fig. 3.

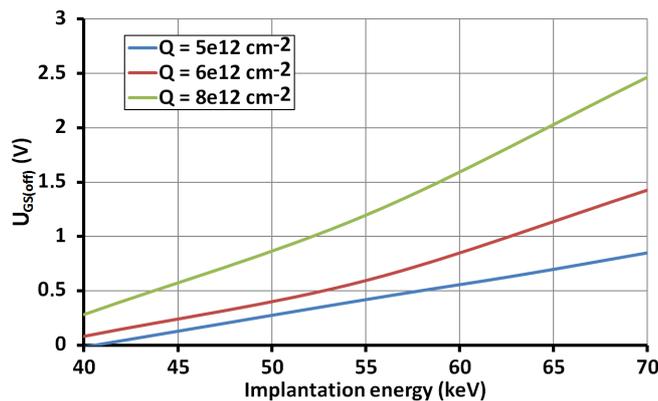


Fig. 2. Dependence of the cutoff voltage of the p-channel JFET on the implantation modes of the collector of the pnp-transistor.

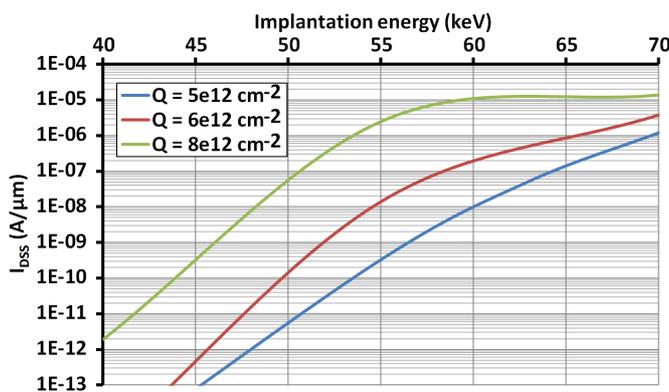


Fig. 3. Dependence of the shorted-gate drain current of the p-channel JFET on the implantation modes of the collector of the pnp-transistor.

Increasing the values of the shorted-gate drain current is possible, first of all, due to increasing the implantation dose. Because of the limitations imposed by the breakdown

voltage of the pnp-transistor, the implantation dose of the channel region cannot be increased by more than $5e12 \text{ cm}^{-2}$. It is possible to increase the drain current values by increasing the implantation energy, which, however, affects the value of the cutoff voltage. For this process, the optimal values of the implantation mode are: energy $E > 70 \text{ keV}$ and dose $Q = 5e12 \text{ cm}^{-2}$.

Further optimization of the parameters of the p-channel field-effect transistor is possible when forming the channel using a separate implantation. This operation should be carried out after the formation of the collector region of the pnp-transistor in order to preserve the gradient impurity profile, which ensures high values of the product $f_T \times U_{CE0}$ of the complementary bipolar transistors.

IV. N-CHANNEL JUNCTION FIELD-EFFECT TRANSISTOR

The design of the integrated n-channel field-effect transistor is shown in Fig. 4.

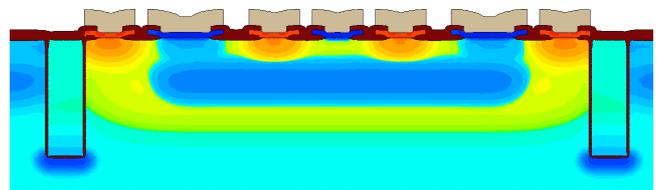


Fig. 4. The design of the integrated n-channel JFET.

The n-channel region of the field-effect transistor is formed by the epitaxial collector of the npn-transistor. At epitaxial film thickness of about 2 microns, the boron diffusion from the buried p+ layer, used as the lower gate, plays a significant role. In this case, a reduction in the channel thickness is observed and high values of the shorted-gate drain current are not provided. The following methods can be used to reduce the back diffusion of boron [15]:

- lowering the temperature of epitaxial growth;
- annealing in an oxidizing atmosphere;
- silicon etching before epitaxy.

Besides, a method of additional (“braking”) doping with phosphorus was proposed, which made it possible to reduce the amount of back diffusion while maintaining the resistance of the region of uniform doping. The advantage of the method is the “selectivity” in the case of using the additional photolithography operation: for the areas of the gate, the diffusion is reduced, while maintaining the low collector resistance of the pnp-transistor. Taking into account the variation of the “braking” doping dose, the possibility of changing the values of the cutoff voltage and the shorted-gate drain current of the n-channel JFET was obtained (Fig. 5).

From the results presented in Fig. 5, it can be seen that the “braking” doping with a dose of $Q = 3e13 \text{ cm}^{-2}$ ensures the symmetry of the values of the cutoff voltage and the shorted-gate drain current with the p-channel junction field-effect transistor described above.

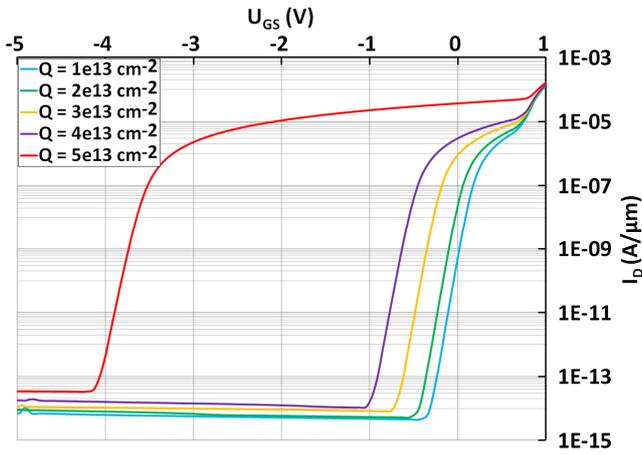


Fig. 5. Dependence of the drain current on the gate-source voltage of n-channel JFET at various doses of "braking" doping ($E = 20 \text{ keV}$).

The results of comparison of the current-voltage characteristics of the complementary junction field-effect transistors are shown in Fig. 6.

As can be seen from Fig. 6, the symmetry of the drain current is observed at the voltages on the gate $U_{GS} = 0 \text{ V}$. With an increase in voltage, the drain current of the n-channel field-effect transistor begins to exceed the drain current of the p-channel transistor. It is worth noting the weak effect of the output voltage on the value of the drain current. The channel length modulation parameter for the n-channel JFET is $\lambda = 0.058 \text{ V}^{-1}$, for the p-channel JFET $\lambda = 0.054 \text{ V}^{-1}$. With an increase in the gate length up to $3 \mu\text{m}$ for both types of transistors, the reduction in the cutoff voltage is $\sim 0.3 \text{ V}$.

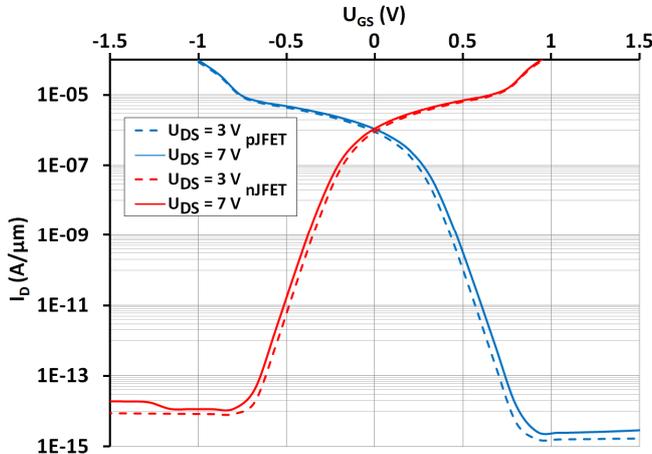


Fig. 6. Dependence of the drain current on the gate-source voltage of the complementary junction field-effect transistors.

V. ADDITIONAL CONSTRUCTIVE FEATURES OF THE COMPLEMENTARY JFETS

To improve the static and dynamic characteristics of the junction field-effect transistors, a design variant was proposed in which the metal drain/source contacts are removed from the gate, and the heavily-doped drain/source regions are close to the gate due to polycrystalline silicon of passive bases of npn- and pnp- transistors. The use of the heavily doped polysilicon and an extended heavily doped area of the drain/source (formed by a sinker of pnp-transistor) for the p-JFET ensured a drain current growth of

more than 25 %. Similar results were obtained for the modified n-channel JFET: the drain current increased by more than 35 %. In this case, the change in the cutoff voltage of the complementary field-effect transistors was no more than 0.5 %.

An important parameter of transistors, significantly dependent on their design and topology, is the breakdown voltage. Reducing the distance between the gate and the heavily-doped drain / source areas at a fixed gate length not only leads to a decrease in resistance, but also to a decrease in the drain-source voltage (U_{DS}). For example, Fig. 7 shows the dependence of U_{DS} of the p-channel field-effect transistor at different distances between the edges of the masks that form the heavily-doped drain/source and gate regions ($W_{G-D/S}$). As can be seen from the figure, with a decrease in the specified distance of less than $1.2 \mu\text{m}$, the drain-source voltage is less than 15 V. At the same time, a significant increase in the modulation parameter of the channel length is observed. Thus, this distance is chosen as the limiting, when optimizing the characteristics of transistors due to modification of the topology. Further improvement of the parameters requires a reduction of process feature size, which will reduce the length of the gate. Table 2 presents the main parameters of the developed complementary junction field-effect transistors.

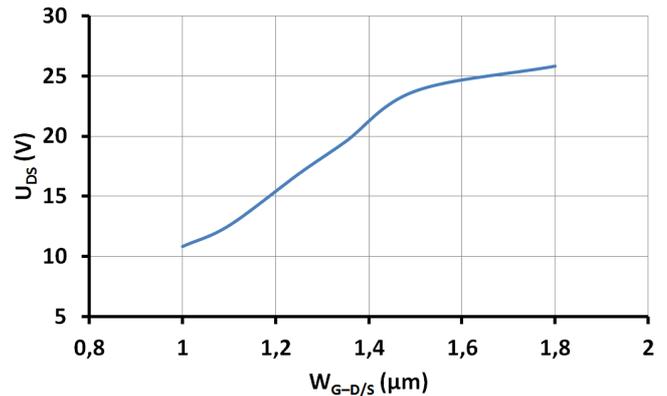


Fig. 7. Dependence of the drain-source voltage of the p-channel junction field-effect transistor on the distance $W_{G-D/S}$.

TABLE II. MAIN PARAMETERS OF THE COMPLEMENTARY JUNCTION FIELD-EFFECT TRANSISTORS

Type of the device	$U_{GS}(\text{off}), \text{ V}$	$f_T, \text{ GHz}$	$I_{DSS}, \text{ A}/\mu\text{m}$	$g_{ms}, \text{ S}/\mu\text{m}$ ($U_{GS} = 0 \text{ V}$)	$U_{DS}, \text{ V}$
	($U_{DS} = 3 \text{ V}$)				
n-JFET	-0.74	0.43	1.04e-6	4.620e-6	>16.5
p-JFET	0.85	0.34	9.36e-7	6.625e-6	>16.7

VI. FEATURES OF THE CIRCUITRY OF THE LOW-TEMPERATURE ANALOG ICs BASED ON THE COMPLEMENTARY FIELD-EFFECT TRANSISTORS

The technological process developed above allows designing analog ICs (AIC), containing both field-effect and microwave bipolar transistors. However, in the range of cryogenic temperatures, the bipolar active elements (especially p-n-p) have extremely small values of the base current gain. Therefore, to build low-temperature and low-noise interfaces of sensors, it is advisable to perform AICs only on CJFet components. In turn, this limitation raises the problem of developing basic CJFet functional units, on the

basis of which various CJFet operational amplifiers, constant voltage regulators, current and voltage comparators, active ARC filters, etc. can be created. The current level of CJFet circuitry is extremely low and not developed. This raises the problem of the CJFet AIC implementation, which is much more complicated than for CMOS and BJ technical processes, which is associated with a different polarity of static voltage between the source and the gate of the junction field-effect transistor in the active mode (compared to the polarity of the drain-source voltage). As a consequence, first of all, we need to search for original CJFet current mirrors (KM) that cannot be performed according to traditional CMOS and BJT schemes of KM, CJFet of input differential stages, buffer amplifiers (BA) with source output, rail-to-rail CJFet BA, input stages of high-speed CJFet op amps, intermediate “bent” cascodes, static mode stabilization circuits (reference current sources and offset potential circuits), high-gain op-amps, high-speed op amps, micropower CJFet op amps, differential difference op amps, OTA-amplifiers with controlled steepness, etc. Essentially, we need an updated design concept for a wide class of analog CJFet microcircuits, which practically did not develop due to the dominant influence of cheap CMOS and BJ technologies, as well as the small amount of low-temperature microelectronic products. However, the modern needs of space instrumentation, high-energy physics, medicine, quantum computers, high-speed rail transport, etc., stimulate the development of this scientific direction. The above circuit problems require urgent solutions.

VII. CONCLUSION

In conclusion, we can formulate the following:

- to ensure symmetric parameters of the complementary junction field-effect transistors within the basis of microwave layers of the complementary bipolar technological process for p-JFETs, it is necessary to use the implantation of the upper region of the p-collector of the pnp-transistor with the energy $E > 70$ keV and the dose $Q = 5e12$ cm⁻²; for n-JFETs it is required to use “braking” doping of suppression of back diffusion of boron with the energy $E = 20$ keV and the dose $Q = 3e13$ cm⁻²;
- the formation mode of the upper p- region is the limit conditions on the minimum value of the breakdown voltage of the pnp-transistor, therefore, that is why, it is advisable to introduce an additional photolithographic operation, which ensures the formation of the p- channel of the junction field-effect transistor;
- the static parameters of the complementary junction field-effect transistors are at the level of the JFET parameters from complementary SiGe bipolar- and JFET-technology BiCom3HV of Texas Instruments [6];
- to further improve the dynamic characteristics of the junction field-effect transistors, the reduction in the gate length is required.

REFERENCES

- [1] G.I. Volovich, Circuitry of the analog and analog-to-digital electronic devices. Dodeka-XXI, 2005 (in Russian).
- [2] O.V. Dvornikov et al., “Cryogenic Operational Amplifier on Complementary JFETs,” 2018 IEEE East-West Design & Test Symposium (EWDTS), 2018, pp. 1-5. DOI: 10.1109/EWDTS.2018.8524640
- [3] P.F. Manfredi, V. Re, V. Speziali, “Monolithic JFET preamplifier with nonresistive charge reset,” IEEE Transactions on nuclear science, 1998, V. 45, No. 4, pp. 2257-2260. DOI: 10.1109/23.709654
- [4] R.T. Goldberg et al., “Fabrication and characterization of low-noise cryogenic Si JFETs,” Proc. Symp. Low Temperature Electronics and High Temperature Superconductivity, 1995, pp. 95-9.
- [5] T. Yang, J. Lu, J. Holleman, “A high input impedance low-noise instrumentation amplifier with JFET input,” 2013 IEEE 56th International Midwest Symposium on Circuits and Systems (MWSCAS), 2013, pp. 173-176. DOI: 10.1109/MWSCAS.2013.6674613
- [6] I.Y. Lovshenko, V.T. Khanko, V.R. Stempitsky, “Radiation influence on electrical characteristics of complementary junction field-effect transistors exploited at low temperatures,” Materials Physics & Mechanics, 2018, V. 39, No. 1.
- [7] W. Schwartz et al., “BiCom3HV-a 36V complementary SiGe bipolar- and JFET-technology,” 2007 IEEE Bipolar/BiCMOS Circuits and Technology Meeting, 2007. DOI: 10.1109/BIPOL.2007.4351835
- [8] M. Snoeij, “A 36V 48MHz JFET-Input Bipolar Operational Amplifier with 150μV Maximum Offset and Overload Supply Current Control,” ESSCIRC 2018-IEEE 44th European Solid State Circuits Conference (ESSCIRC), IEEE, 2018, pp. 290-293. DOI: 10.1109/ESSCIRC.2018.8494262
- [9] M. Dentan et al., “Study of a CMOS-JFET-bipolar radiation hard analog-digital technology suitable for high energy physics electronics,” IEEE transactions on nuclear science, 1993, V. 40, No. 6, pp. 1555-1560. DOI: 10.1109/23.273505
- [10] M. Manghisoni et al., “Selection criteria for P-and N-channel JFETs as input elements in low-noise radiation-hard charge preamplifiers,” IEEE Transactions on Nuclear Science, 2001, V. 48, No. 4, pp. 1598-1604. DOI: 10.1109/23.958402
- [11] D.G. Drozdov, E.M. Savchenko, “Features of the self-aligning technology when designing the complementary bipolar transistors,” Electronics. Series 2 Semiconductor devices, 2011, No.2, pp. 53–58 (in Russian).
- [12] D.G. Drozdov, E.M. Savchenko, A.M. Zubkov, “The results of the instrumental-process simulation of the complementary bipolar technology with the cutoff frequency of 10 GHz and more,” Problems of the development of the advanced micro- and nano-electronic systems - 2010. Collection of research papers / under general editorship of the academician A.L. Stempkovsky. M.:IPPM RAS, 2010, pp. 66-69 (in Russian).
- [13] R.N. Vinogradov, D.G. Drozdov, P.A. Dukanov, D.L. Ksenofontov, S.V. Korneev, E.M. Savchenko, G.P. Surkov, “Optimization of the technological process of creation of ICs of the integrated junction field-effect transistors,” Solid-state electronics. Complex operating assemblies of Electronics. Materials of the XI-th All-Russian scientific and technical conference. Moscow: MNTORES n.a. A.S. Popova, 2012, pp. 222–223 (in Russian).
- [14] E.M. Savchenko, D.G. Drozdov, A.V. Vagin, D.I. Garanovich, “Modern constructions of the integrated elements of the high-frequency complementary bipolar technological process,” Fundamental problems of radioelectronic tool engineering. Materials of the International scientific and technical conference “INTERMATIC-2013”. Part 3. Moscow: Energoatomizdat, 2013, pp. 60-63 (in Russian).
- [15] D.G. Drozdov, “Microwave complementary bipolar technological process with high degree of symmetry of dynamic parameters of the transistors,” Ph.D. thesis in Engineering Science. Speciality 05.27.01 – Solid-state electronics, radioelectronic components, micro- and nano-electronics, devices on quantum effects. Moscow, 2017, pp. 165 (in Russian).

The Discrete Structure of the Zeros and Poles Location in the z-Plane of the Arbitrary Order IIR Digital Filters with a Finite Word Length

Vladislav Lesnikov¹, Tatiana Naumovich², Alexander Chastikov³, Alexander Metelyov⁴
Department of radioelectronic systems,
Vyatka State University
Kirov, Russia

¹vladislav.lesnikov.ru@ieec.org, ²ntv_new@mail.ru, ³alchast@mail.ru, ⁴metap@inbox.ru

Abstract— It is known that the resolved positions of the zeros and poles of the second order IIR digital filters with finite wordlength form a discrete structure called in this paper the topography of a discretized z-plane. In the publications of the authors, the z-plane topography was described in detail. This paper is devoted to the definition of equations describing plane algebraic curves on which the allowed positions for the zeros and poles of filters of arbitrary order are located.

Keywords— IIR digital filters, Possible pole-zero locations, Grid of allowable pole and zero positions, Variation curve of the poles and zeros, Quantization of pole locations, Plane algebraic curves

I. INTRODUCTION

Despite the fact that the theory of designing IIR digital filters has been developed for a long time, the difficulties accompanying the practical development process in the case of stringent requirements for characteristics turn out to be so complex that developers have to abandon IIR filters in favor of FIR filters. Overcoming the problems arising in this case requires a deep study of the nature of IIR filters with a finite word length (FWL).

The traditional approach to the synthesis of IIR digital filters [1] includes the stages of functional and structural synthesis. Functional synthesis involves the calculation of the zeros and poles of the transfer function without taking into account the FWL. The finite bit depth of the filter coefficients is taken into account only at the stage of structural synthesis. Therefore, the results of structural synthesis distort the results of functional synthesis and, therefore, errors are introduced into the filter characteristics. Accounting for these circumstances complicates the procedures necessary to meet the requirements for filter characteristics.

In [2] we are developing a new approach to the synthesis of IIR digital filters with FWL. This approach includes the requirement to take into account FWL already at the stage of functional synthesis. At this stage, the zeros and poles are finally calculated and not distorted during structural synthesis. At the stage of structural synthesis, the structure is generated.

The theoretical basis of our approach to functional synthesis is based on the fact that the roots of the polynomials of the numerator and denominator of the transfer function of the practical implemented digital filters are elements of the set of algebraic numbers [3]. This means that not every point in the z-plane can be a zero or a pole of a FWL digital filter. The allowed positions of zeros and poles form a discrete structure in the z-plane, which we call topography.

For the poles of FWL second-order digital filters, this has long been known [4] - [6]. This topography was studied in detail in [7], [8]. For a long time, we tried to extend this theory to IIR FWL filters of higher orders. Only recently managed to get results for filters of the third [9] and fourth [10] order.

In this paper, we propose a technique that allows one to describe the topography of a discretized z-plane for algebraic numbers of arbitrary degree.

II. THE ALGEBRAIC-NUMERIC NATURE OF THE ZEROS AND POLES OF FWL DIGITAL FILTERS

Information from the Theory of Algebraic Numbers

It is known [11] that the roots $z_{n,i}$ of polynomial

$$P_n(z) = \sum_{i=0}^n c_{n,i} z^{n-i} \quad (1)$$

with real rational coefficients

$$c_{n,i} \in \mathbb{Q} \quad (2)$$

are elements of the set of complex algebraic numbers

$$z_{n,i} \in \mathbb{A}_n \subset \mathbb{A}, \quad (3)$$

where n is the algebraic number degree.

It is obvious that the coefficients of practically implemented digital filters have a finite bitness and, therefore, are rational numbers. The coefficients $b_{n,i}$ and $a_{n,i}$ of the transfer function

The work was supported by a grant from the Russian Foundation for Basic Research 18-07-00986.

$$\begin{cases} c_{n,1} & = & -2x & +c_{n-2,1}, \\ c_{n,2} & = & (x^2 + y^2) & -2xc_{n-2,1} +c_{n-2,2}, \\ c_{n,3} & = & (x^2 + y^2)c_{n-2,1} & -2xc_{n-2,2} +c_{n-2,3}, \\ \dots & \dots & \dots & \dots \\ c_{n,i} & = & (x^2 + y^2)c_{n-2,i-2} & -2xc_{n-2,i-1} +c_{n-2,i}, \\ \dots & \dots & \dots & \dots \\ c_{n,n-2} & = & (x^2 + y^2)c_{n-2,n-4} & -2xc_{n-2,n-3} +c_{n-2,n-2}, \\ c_{n,n-1} & = & (x^2 + y^2)c_{n-2,n-3} & -2xc_{n-2,n-2}, \\ c_{n,n} & = & (x^2 + y^2)c_{n-2,n-2}. \end{cases} \quad (15)$$

We will consider (15) as a system of equations from which it is necessary to obtain the function $y^2(x)$. The parameters of this function should be the coefficients of the polynomial $P_n(z)$, and the coefficients of the polynomial $P_{n-2}(z)$ should be excluded. Therefore, (15) must be solved with respect to y and the coefficients $c_{n-2,1}, c_{n-2,2}, \dots, c_{n-2,n-2}$. In order for the number of equations to be equal to the number of unknown quantities, we will assume that one of the coefficients $c_{n,i}$ is unknown.

Thus, from (15) the following functions

$$y^2 = y_{n,i}^2(x | \mathbf{c}_{n,i}), \quad (16)$$

can be obtained. The set \mathbf{c}_i is defined as the set of coefficients of the polynomial $P_n(z)$, from which the coefficient $c_{n,i}$ is excluded. Given the values of the quantized coefficients in $\mathbf{c}_{n,i}$, we obtain a family of curves on which the allowed values of the roots of the polynomial P are located. Any curve (16) is a geometrical place of all allowed roots. To solve the equations, the capabilities of the Maple (MapleSoft) system to perform symbolic calculations will be used.

V. THE TOPOGRAPHY OF THE THIRD-DEGREE ALGEBRAIC NUMBERS

For the polynomial $P_3(z)$ (15) is converted to

$$\begin{cases} c_{3,1} & = & -2x & +c_{1,1}, \\ c_{3,2} & = & (x^2 + y^2) & -2xc_{1,1}, \\ c_{3,3} & = & (x^2 + y^2)c_{1,1}. \end{cases} \quad (17)$$

From (17) we obtain solutions

$$y_{3,1}^2(x | \mathbf{c}_{3,1}) : \left\{ y^2 = 0.5c_{3,2} \pm 0.5\sqrt{c_{3,2}^2 + 8c_{3,3}x - x^2} \right\}, \quad (18)$$

$$y_{3,2}^2(x | \mathbf{c}_{3,2}) : \left\{ y^2 = -\frac{2x^3 + c_{3,1}x^2 - c_{3,3}}{c_{3,1} + 2x} \right\}, \quad (19)$$

$$y_{3,3}^2(x | \mathbf{c}_{3,3}) : \left\{ y^2 = 3x^2 + 2c_{3,1}x + c_{3,2} \right\}. \quad (20)$$

Simple transformations (18) - (19) lead to the equations

$$Y_{3,1}(x | \mathbf{c}_{3,1}) : \left\{ (x^2 + y^2)^2 - c_{3,2}(x^2 + y^2) - 2c_{3,3}x = 0 \right\}, \quad (21)$$

$$Y_{3,2}(x | \mathbf{c}_{3,2}) : \left\{ 2x^3 + 2xy^2 + c_{3,1}(x^2 + y^2) - c_{3,3} = 0 \right\}, \quad (22)$$

$$Y_{3,3}(x | \mathbf{c}_{3,3}) : \left\{ y^2 - 3x^2 - 2c_{3,1}x - c_{3,2} = 0 \right\}. \quad (23)$$

The equations $Y_i(x | \mathbf{c}_{3,i})$ ($i=1,2,3$) describe, respectively, the plane algebraic curves of the fourth (quartics), third (cubics), and second (conics) order. Each of these equations contains two parameters. For convenience of visualization, we will depict the curve $Y_i(x | \mathbf{c}_{3,i})$ in such a way that each image will correspond to a section of the space of coefficients by a plane parallel to one of the coordinate planes. In this case, we obtain a family of curves determined by one parameter (Fig. 2).

VI. THE TOPOGRAPHY OF THE FOURTH-DEGREE ALGEBRAIC NUMBERS

For polynomial $P_4(z)$, we obtain the system of equations:

$$\begin{cases} c_{4,1} & = & -2x & +c_{2,1}, \\ c_{4,2} & = & (x^2 + y^2) & -2xc_{2,1} +c_{2,2}, \\ c_{4,3} & = & (x^2 + y^2)c_{2,1} & -2xc_{2,2}, \\ c_{4,4} & = & (x^2 + y^2)c_{2,2}. \end{cases} \quad (23)$$

The solution for $y_{4,1}^2(x | \mathbf{c}_{4,1})$ is obtained in the form:

$$\widehat{Y}_1(x | \mathbf{c}_{4,1}) : \left\{ Y = \frac{c_{4,4}}{\text{RootOf}(X^2Z^3 + (-Xc_{4,3} - c_{4,4})Z^2 + c_{4,2}c_{4,4}Z - c_{4,4}^2)} \right\}, \quad (24)$$

where

$$\begin{cases} X = -2x, \\ Y = x^2 + y^2, \end{cases} \quad (25)$$

the function of Maple RootOf is a placeholder for representing all the roots of an equation in one variable. In this case, RootOf(F(Z)) describes the roots of the equation

$$F(Z) = 0. \quad (26)$$

From equation (24) it follows that

$$Z = \frac{c_{4,4}}{Y} \quad (27)$$

is the root of the equation that describes the desired curve. After simple transformations we get

$$Y_{4,1}(x | \mathbf{c}_{4,1}) : \left\{ -(x^2 + y^2)^3 + (x^2 + y^2)^2 c_{4,2} + (x^2 + y^2)(2x - c_{4,4}) + 4x^2 c_{4,4} = 0 \right\}. \quad (28)$$

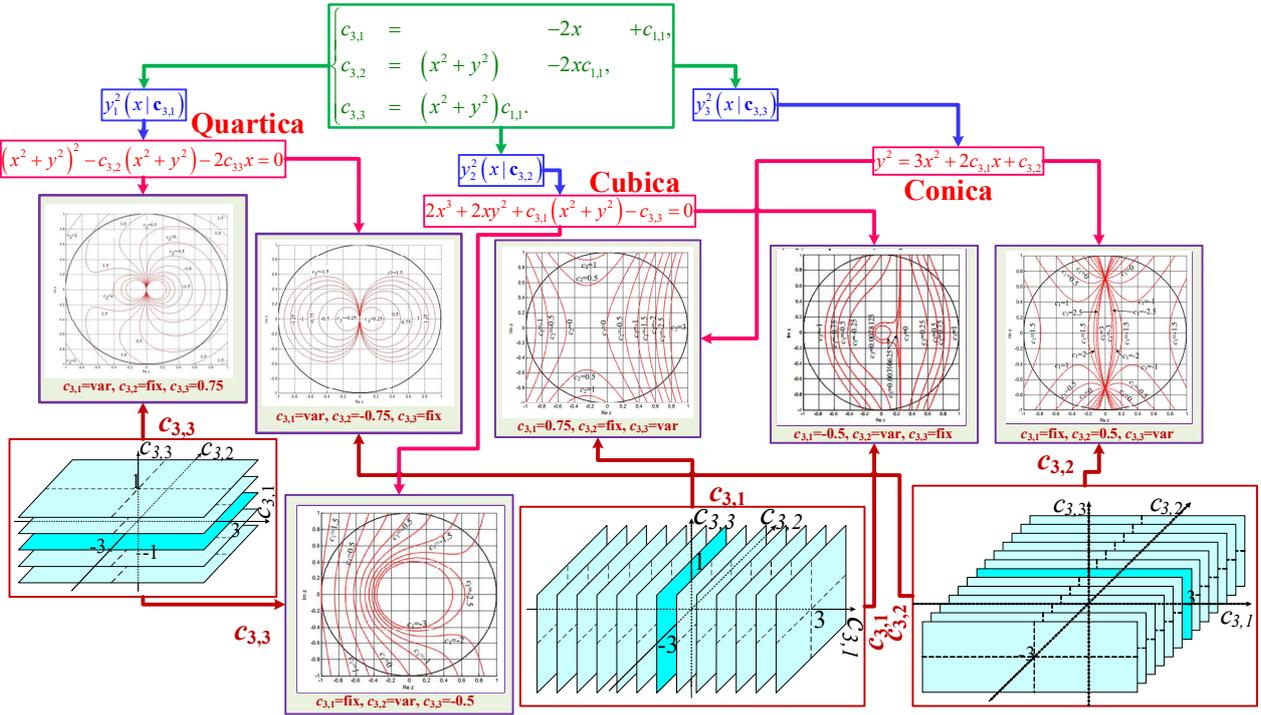


Fig. 2. Calculation of the geometric place of third-degree algebraic numbers in various ways in different sections of the polynomial coefficient space.

The intermediate representation of the remaining equations is:

$$\hat{Y}_{4,2}(x | \mathbf{c}_{4,2}): \{Y = \text{RootOf}((-c_{41} + X)Z^2 + c_{43}Z - c_{44}X)\}, \quad (29)$$

$$\hat{Y}_{4,3}(x | \mathbf{c}_{4,3}): \{Y = X^2 - Xc_{41} - \text{RootOf}(Z^2 + (-X^2 + Xc_{41} - c_{42})Z + c_{44}) + c_{42}\}, \quad (30)$$

$$\hat{Y}_{4,4}(x | \mathbf{c}_{4,4}): \left\{Y = \frac{X^3 - X^2c_{41} + Xc_{42} - c_{43}}{2X - c_{41}}\right\}. \quad (31)$$

The final form of the equations is:

$$Y_{4,2}(x | \mathbf{c}_{4,2}): \{(-c_{4,1} - 2x)(x^2 + y^2)^2 + c_{4,3}(x^2 + y^2) + 2c_{4,4}x = 0\}, \quad (32)$$

$$Y_{4,3}(x | \mathbf{c}_{4,3}): \{(x^2 + y^2)^2 - (x^2 + y^2)(4x^2 + 2xc_{4,1} + c_{4,2}) + c_{4,4} = 0\}, \quad (33)$$

$$Y_{4,4}(x | \mathbf{c}_{4,4}): \{8x^3 + (x^2 + y^2)(-4x - c_{4,1}) + 4x^2c_{4,1} + 2xc_{4,2} + c_{4,3} = 0\}. \quad (34)$$

VII. THE TOPOGRAPHY OF THE HIGHER-DEGREE ALGEBRAIC NUMBERS

This section presents the results of the derivation of implicit equations describing plane algebraic curves, on which the resolved positions of the zeros and poles of the IIR FWL digital filters are located. The section has a reference character. Only final results are presented. The derivation of equations is omitted. The inference technique is similar to the one that was given in the previous sections.

A. The Fifth-Degree Algebraic Numbers

The final solutions:

$$Y_{5,1}(x | \mathbf{c}_{5,1}): \left\{ \begin{aligned} &(x^2 + y^2)^4 - c_{5,2}(x^2 + y^2)^3 + (-2c_{5,3}x + c_{5,4})(x^2 + y^2)^2 + \\ &+ (-4c_{5,4}x^2 + 4c_{5,5}x)(x^2 + y^2) - 8x^3c_{5,5} = 0, \end{aligned} \right\} \quad (35)$$

$$Y_{5,2}(x | \mathbf{c}_{5,2}): \left\{ \begin{aligned} &(-2x - c_{5,1})(x^2 + y^2)^3 + c_{5,3}(x^2 + y^2)^2 + \\ &+ (2c_{5,4}x - c_{5,5})(x^2 + y^2) + 4x^2c_{5,5} = 0, \end{aligned} \right\} \quad (36)$$

$$Y_{5,3}(x | \mathbf{c}_{5,3}): \left\{ \begin{aligned} &(x^2 + y^2)^3 + (-2c_{5,1}x - 4x^2 - c_{5,2})(x^2 + y^2)^2 + \\ &+ c_{5,4}(x^2 + y^2) + 2xc_{5,5} = 0 \end{aligned} \right\} \quad (37)$$

$$Y_{5,4}(x | \mathbf{c}_{5,4}): \left\{ \begin{aligned} &(-4x - c_{5,1})(x^2 + y^2)^2 + \\ &+ (4c_{5,1}x^2 + 8x^3 + 2c_{5,2}x + c_{5,3})(x^2 + y^2) - c_{5,5} = 0 \end{aligned} \right\} \quad (38)$$

$$Y_{5,5}(x | \mathbf{c}_{5,5}): \left\{ \begin{aligned} &(x^2 + y^2)^2 + (-4c_{5,1}x - 12x^2 - c_{5,2})(x^2 + y^2) + \\ &+ 16x^4 + 8x^3c_{5,1} + 4x^2c_{5,2} + 2xc_{5,3} + c_{5,4} = 0 \end{aligned} \right\} \quad (39)$$

B. The Sixth-Degree Algebraic Numbers

The final solutions:

$$Y_{6,1}(x | \mathbf{c}_{6,1}): \left. \begin{aligned} &(x^2 + y^2)^5 - c_{6,2}(x^2 + y^2)^4 + \\ &+ (-2c_{6,3}x + c_{6,4})(x^2 + y^2)^3 + \\ &+ (-4c_{6,4}x^2 + 4c_{6,5}x - c_{6,6})(x^2 + y^2)^2 + \\ &+ (-8c_{6,5}x^3 + 12c_{6,6}x^2)(x^2 + y^2) - 16x^4c_{6,6} = 0 \end{aligned} \right\}, \quad (40)$$

$$Y_{6,2}(x | \mathbf{c}_{6,2}): \left. \begin{aligned} &(-c_{6,1} - 2x)(x^2 + y^2)^4 + c_{6,3}(x^2 + y^2)^3 + \\ &+ (2c_{6,4}x - c_{6,5})(x^2 + y^2)^2 + \\ &+ (4c_{6,5}x^2 - 4c_{6,6}x)(x^2 + y^2) + 8x^3c_{6,6} = 0 \end{aligned} \right\}, \quad (41)$$

$$Y_{6,3}(x | \mathbf{c}_{6,3}): \left. \begin{aligned} &(x^2 + y^2)^4 + (-2c_{6,1}x - 4x^2 - c_{6,2})(x^2 + y^2)^3 + \\ &+ c_{6,4}(x^2 + y^2)^2 + (2c_{6,5}x - c_{6,6})(x^2 + y^2) + 4x^2c_{6,6} = 0 \end{aligned} \right\} \dots (42)$$

$$Y_{6,4}(x | \mathbf{c}_{6,4}): \left. \begin{aligned} &(-4x - c_{6,1})(x^2 + y^2)^3 + \\ &+ (4c_{6,1}x^2 + 8x^3 + 2c_{6,2}x + c_{6,3})(x^2 + y^2)^2 - \\ &- c_{6,5}(x^2 + y^2) - 2xc_{6,6} = 0 \end{aligned} \right\}, \quad (43)$$

$$Y_{6,5}(x | \mathbf{c}_{6,5}): \left. \begin{aligned} &(x^2 + y^2)^3 + (-4c_{6,1}x - 12x^2 - c_{6,2})(x^2 + y^2)^2 + \\ &+ (8c_{6,1}x^3 + 16x^4 + 4c_{6,2}x^2 + 2c_{6,3}x + c_{6,4})(x^2 + y^2) - \\ &- c_{6,6} = 0 \end{aligned} \right\}, \quad (44)$$

$$Y_{6,6}(x | \mathbf{c}_{6,6}): \left. \begin{aligned} &(-6x - c_{6,1})(x^2 + y^2)^2 + \\ &+ (12c_{6,1}x^2 + 32x^3 + 4c_{6,2}x + c_{6,3})(x^2 + y^2) - \\ &- 32x^5 - 16x^4c_{6,1} - 8x^3c_{6,2} - 4x^2c_{6,3} - 2xc_{6,4} - c_{6,5} = 0 \end{aligned} \right\}. \quad (45)$$

C. The Sixth-Degree Algebraic Numbers

The final solutions:

$$Y_{7,1}(x | \mathbf{c}_{7,1}): \left. \begin{aligned} &(x^2 + y^2)^6 - c_{7,2}(x^2 + y^2)^5 + \\ &+ (-2c_{7,3}x + c_{7,4})(x^2 + y^2)^4 + \\ &+ (-4c_{7,4}x^2 + 4c_{7,5}x - c_{7,6})(x^2 + y^2)^3 + \\ &+ (-8c_{7,5}x^3 + 12c_{7,6}x^2 - 6c_{7,7}x)(x^2 + y^2)^2 + \\ &+ (-16c_{7,6}x^4 + 32c_{7,7}x^3)(x^2 + y^2) - 32x^5c_{7,7} = 0 \end{aligned} \right\}, \quad (46)$$

$$Y_{7,2}(x | \mathbf{c}_{7,2}): \left. \begin{aligned} &(-c_{7,1} - 2x)(x^2 + y^2)^5 + c_{7,3}(x^2 + y^2)^4 + \\ &+ (2c_{7,4}x - c_{7,5})(x^2 + y^2)^3 + \\ &+ (4c_{7,5}x^2 - 4c_{7,6}x + c_{7,7})(x^2 + y^2)^2 + \\ &+ (8c_{7,6}x^3 - 12c_{7,7}x^2)(x^2 + y^2) + 16x^4c_{7,7} = 0 \end{aligned} \right\}, \quad (47)$$

$$Y_{7,3}(x | \mathbf{c}_{7,3}): \left. \begin{aligned} &(x^2 + y^2)^5 + (-2c_{7,1}x - 4x^2 - c_{7,2})(x^2 + y^2)^4 + \\ &+ c_{7,4}(x^2 + y^2)^3 + (2c_{7,5}x - c_{7,6})(x^2 + y^2)^2 + \\ &+ (4c_{7,6}x^2 - 4c_{7,7}x)(x^2 + y^2) + 8x^3c_{7,7} = 0 \end{aligned} \right\}, \quad (48)$$

$$Y_{7,4}(x | \mathbf{c}_{7,4}): \left. \begin{aligned} &(-4x - c_{7,1})(x^2 + y^2)^4 + \\ &+ (4c_{7,1}x^2 + 8x^3 + 2c_{7,2}x + c_{7,3})(x^2 + y^2)^3 - \\ &- c_{7,5}(x^2 + y^2)^2 + \\ &+ (-2c_{7,6}x + c_{7,7})(x^2 + y^2) - 4x^2c_{7,7} = 0 \end{aligned} \right\}, \quad (49)$$

$$Y_{7,5}(x | \mathbf{c}_{7,5}): \left. \begin{aligned} &(x^2 + y^2)^4 + (-4c_{7,1}x - 12x^2 - c_{7,2})(x^2 + y^2)^3 + \\ &+ (8c_{7,1}x^3 + 16x^4 + 4c_{7,2}x^2 + 2c_{7,3}x + c_{7,4})(x^2 + y^2)^2 - \\ &- c_{7,6}(x^2 + y^2) - 2xc_{7,7} = 0 \end{aligned} \right\}, \quad (50)$$

$$Y_{7,6}(x | \mathbf{c}_{7,6}): \left. \begin{aligned} &(-6x - c_{7,1})(x^2 + y^2)^3 + \\ &+ (12c_{7,1}x^2 + 32x^3 + 4c_{7,2}x + c_{7,3})(x^2 + y^2)^2 + \\ &- (16c_{7,1}x^4 + 32x^5 + 8c_{7,2}x^3 + 4c_{7,3}x^2 + 2c_{7,4}x + c_{7,5})(x^2 + y^2) + \\ &+ c_{7,7} = 0 \end{aligned} \right\}, \quad (51)$$

$$Y_{7,7}(x | \mathbf{c}_{7,7}): \left. \begin{aligned} &(x^2 + y^2)^3 + (-6c_{7,1}x - 24x^2 - c_{7,2})(x^2 + y^2)^2 + \\ &+ (32c_{7,1}x^3 + 80x^4 + 12c_{7,2}x^2 + 4c_{7,3}x + c_{7,4})(x^2 + y^2) - \\ &- 64x^6 - 32x^5c_{7,1} - 16x^4c_{7,2} - 8x^3c_{7,3} - \\ &- 4x^2c_{7,4} - 2xc_{7,5} - c_{7,6} = 0 \end{aligned} \right\}. \quad (52)$$

CONCLUSIONS

In this paper, a system of equations is obtained, the solution of which allows us to obtain equations of plane algebraic curves, on which the positions of all possible zeros and poles of the IIR FWL digital filters of arbitrary order are located. The solution of the derived system of equations should be carried out by the methods of symbolic mathematics. This article uses the capabilities of the Maple computer math system. The equations of plane algebraic curves obtained for filters up to the seventh order are presented. The results of the work will be

used in the implementation of the approach developed by the authors to the synthesis of IIR FWL filters, in which the final word length is taken into account when calculating zeros and poles even before the stage of structural synthesis. In this case, structural synthesis does not distort the calculated values of zeros and poles.

REFERENCES

- [1] D. Schlichthärle, *Digital Filters: Basics and Design*, Berlin, Heidelberg: Springer-Verlag, 2011.
- [2] V. Lesnikov, A. Chastikov, T. Naumovich, and S. Armishev, "A new paradigm in design of IIR digital filters," *8th IEEE East-West Design and Test Symposium (EWDTS 2010)*, St. Petersburg, Russia, 17-20 Sept. 2010, pp. 282-285.
- [3] V. Lesnikov, T. Naumovich, and A. Chastikov, "Number-theoretical analysis of the structures of classical IIR digital filters," *7th Mediterranean Conference on Embedded Computing (MECO 2018)*, Budva, Montenegro, 10-14 June 2018, 4 p., <https://doi.org/10.1109/MECO.2018.8406099>.
- [4] C. J. Weinstein, *Quantization Effects in Digital Filters*, Technical Report 468, Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts, 21 November 1969, available: <https://www.semanticscholar.org/paper/Quantization-Effects-in-Digital-Filters-Weinstein/0e52d6dbb14fb6c137527f7919e0bc380bd276f8>.
- [5] W. Hess, *Digitale Filter: eine Einführung*, Springer Fachmedien Wiesbaden GmbH, 1993.
- [6] B. W. Bomar, "Finite Wordlength Effects," in *Digital Signal Processing Handbook*, ed. V. K. Madisetti, and D. B. Williams, Boca Raton: CRC Press LLC, 1999.
- [7] V. Lesnikov, T. Naumovich, and A. Chastikov, "Topography of z-plane which is discretized due to quantization of coefficients of digital biquad filters," *12th International Siberian Conference on Control and Communications (SIBCON 2016)*, Moscow, Russia, 12-14 May 2016, 4 p., <https://doi.org/10.1109/SIBCON.2016.7491812>.
- [8] V. Lesnikov, T. Naumovich, and A. Chastikov, "The sampling of the z-plane due to the quantization of the digital filter coefficients," *7th Mediterranean Conference on Embedded Computing (MECO 2018)*, Budva, Montenegro, 10-14 June 2018, 4 p. <https://doi.org/10.1109/MECO.2018.8405962>.
- [9] V. Lesnikov, T. Naumovich, and A. Chastikov, "The topography of a third order IIR digital filter zeros and poles in the z-plane discretized due to the quantization of the direct form coefficients," *7th Mediterranean Conference on Embedded Computing (MECO 2019)*, Budva, Montenegro, 10-14 June 2019, pp. 374-377.
- [10] V. Lesnikov, T. Naumovich, A. Chastikov, and A. Metelyov, "Topography of the z-plane discretized by quantizing the coefficients of the canonical form of recursive digital filter," in *Computer Vision in Advanced Control Systems - 6*, M. Favorskaya, and L.C. Jain, Eds, in press (will be published by Springer in 2020)
- [11] D. Hilbert, *The Theory of Algebraic Number Fields*, Berlin – Heidelberg – New York: Springer – Verlag, 1998..

Permanent Monitoring Systems of the Contact-Wire of Railroad Catenary: the Main Tasks of Implementation

Dmitrii V. Efanov,
DSc, Professor at "Automation,
Remote Control and
Communication on Railway
Transport", Russian University
of Transport (MIIT),
Moscow, Russia
TrES-4b@yandex.ru

German V. Osadchy,
Technical Director of Scientific
and Technical Center
"Integrated Monitoring
Systems" LLC,
St. Petersburg, Russia
osgerman@mail.ru

Dmitrii V. Barch,
Head of traction power
engineering laboratory
of the October Directorate
of Energy Supply
at JCS "Russian Railways",
St. Petersburg, Russia
barchdv@rambler.ru

Andrei A. Belyi,
PhD, associate professor,
cathedra "Bridges" chief,
Emperor Alexander I
St. Petersburg State Transport
University,
St. Petersburg, Russia
andbelyi@mail.ru

Abstract—Technical condition of major and auxiliary sites of transportation infrastructure, including railroad branch, must be determined on time and hidden defects should not cause to failures which impede technological schedule fulfillment. Well timed definition of sophisticated railroad structural technical elements status quo is impossible to complete by means of maintenance specialists only, but permanent monitoring performance is essential to achieve the required task. The aforesaid is the issue for the railroad catenary as well. Permanent monitoring of supporting structural elements of railroad catenary is being conducted via railway-laboratory vehicles as well as by means of maintenance workers in accordance with available manuals and instructions. Well known permanent monitoring systems are being applied within many countries. As for the Russian Federation, for the term of permanent monitoring stage the above systems were not being implemented, but experimental researches are being conducted. It is worthy of note that authors of present paper were pioneers of those systems for Russian Railroad Network. Our system is considered as universal one and may be suitable per any type of railroad catenary including tackle gear compensation devices. Vast experience of authors concerning design and services of the above-mentioned systems as well as catenary kit itself, helps us to take a look in nearest future of permanent monitoring systems. Our present paper is aimed at this important factor highlighted.

Keywords: *railroad catenary, reliability, catenary suspension, technical maintenance, monitoring, pre-failure condition*

I. INTRODUCTION

The Russian Federation possesses world vast net of electrified railroads with the entire length of forty thousand kilometers. The only one country with the same railroad features to compare is China.

The quality of railroad catenary auxiliary elements is the function of steady railroad traffic performance with proper train schedule fulfillment.

Compared to, for instance, most parts of railroad catenary automation are being installed without any backup [1]. Railroad catenary belongs to the aforesaid elements being in contact with pantograph of rolling stock. More than that, the condition of railroad catenary is the function of rolling stock automation status quo (for example, track circuit failure to-

gether with locomotive signalling, as well as collector bow damage during interaction with railroad catenary). Statistics shows, that the most unreliable elements of railroad catenary structure are cables with much more failures quantity than other structural essentials (subsequently the list of unreliable elements shows strings, clamps, railroad air switches) [2]. To ensure the better reliability and safety of railroad catenary the set of technical diagnostic procedures and monitoring are being performed. Those measures are being conducted by railroad service crews as well as by special railway-laboratory vehicles aimed at catenary status [3]. Recently new permanent monitoring systems are the matter of fact reckoning various characteristics of railway catenary. Those approaches becoming more and more popular abroad plus helps us to improve the service of railroad catenary [4 – 13].

One of the first permanent monitoring system aimed at specific of railroad catenary and being functioned on Post Soviet terrain is showed in [14]. The above system was arranged by means of cooperation with present authors and initially was designed on cable vibration analysis of railroad catenary. Our experience concerning permanent monitoring services of high speed railroad Moscow – St. Petersburg highlights insufficiency of such measurements for definitions of pre-failure status of railroad auxiliary structural elements. Our engineers did upgrade the system via application of not vibration control only, but by means of measurement of mechanical impacts on cables [15], plus piers inclination tendency supervision [16, 17]. Modernized system is being tested for further integration with railroad catenary. It ought to be remarked the importance of such researches with example of high speed railroad line Moscow – Kazan, where was conceived technical decision of diagnostic gadgets installation for permanent monitoring system of railroad catenary kit. Based on analytical issue, permanent monitoring systems of railroad await to be in progress in nearest future. In present contribution we pay attention on promising trends of railroad catenaries supervision in accordance with years of experience regarding its service and maintenance. The article purpose is to highlight the most important stages in the monitoring system implementation of the railway contact-wire of catenary system in the post-Soviet space. Based on the operating experience of the first versions of the contact suspension monitoring system [14 – 16], the authors

presented the necessary set of diagnostic parameters for organizing a qualitative analysis of this diagnostic object. The most rational places for connecting diagnostic devices are indicated, considering the types of contact suspension and economic considerations of the cost of the monitoring system itself. Recommendations for the development of a monitoring system for contact suspension for post-Soviet railroads are formulated, taking into account their specificity.

II. DEVELOPMENT TRENDS OF PERMANENT MONITORING SYSTEMS

The most frequent and harmful damages of railroad catenary elements are considered wires plus cables, so this is the root matter to keep your mind on.

For better status quo definition of railroad catenary parts we propose to control the features of strands vibration and tension [14, 15]. Oscillation supervision helps us to identify such troubling events as: tendons/cables breakage, cable messenger damage, impacts of pantograph etc. Cables tension monitoring allows us to define the condition of counter-balance weights (hoisting tackle wedging, kentledge theft, non-normatively of loads position etc.), tension control should help us to clarify the condition of cantilevers as well (wedging event) by means of middle anchorage spot supervision. It should be noted that stressing of railroad catenary is very sensitive issue regarding proper system functioning.

During low temperature events catenary breakage factor takes places more frequently consequently, sensors installation within tendons (A-185, AC-185, M-120) considered as an crucial precaution. Worth-while to clarify those events of wires damages for decision making reckoning spots per vibration sensors assembling. Places of installation must be specified in case of two or three tendons layout.

Suggested schema of sensors arrangement per elements of railroad catenary presented on Fig. 1.

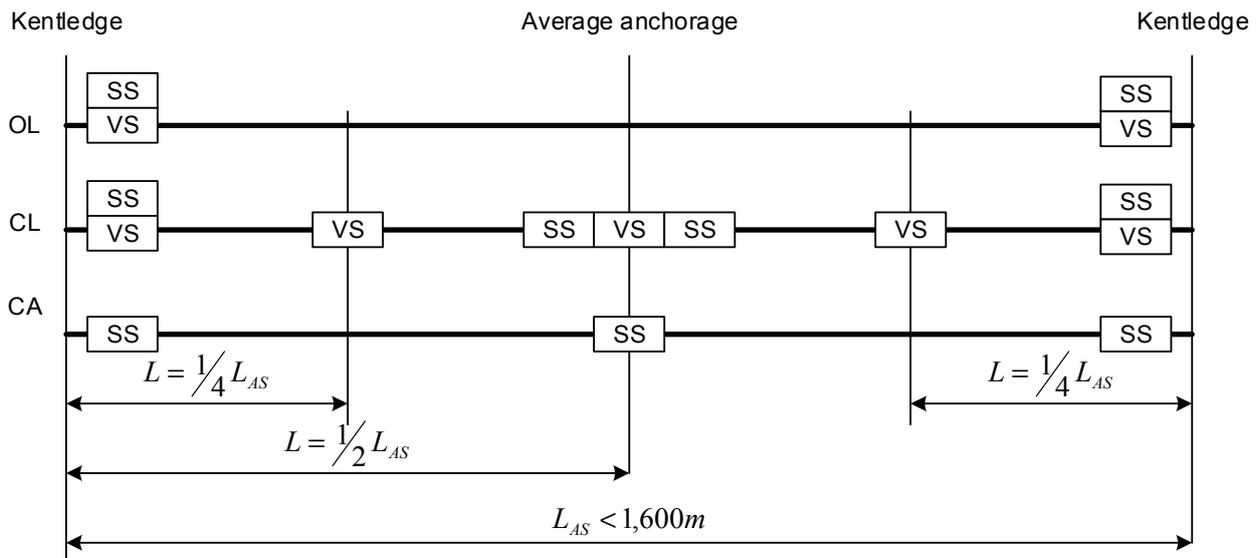
One more important item of the catenary kit is those cable supporting masts structures [16]. In case of supporting pole fall down trouble, rehabilitation time sector is much longer compares to other possible reimbursement at railroad site (except flexible and rigid beams). As for price per single mast exchange, based on records, it was about \$6,000 back in 2016. Preferable to supervise oscillation features with angle of vibration per piers, for the reason of the above characteristics alteration considered as a pier/foundation structural status quo deterioration [18]. Within areas of direct current (DC) performance regarding railroad catenary, current leakage shall be the matter concerning locomotive automation signalling.

Besides, monitoring procedure should be conducted within stations reckoning pantograph with catenary tension via distance definition between basic and auxiliary catenary clips during trains passage moment (for scientific research purposes we may apply 'wave' features per high speed areas of transportation).

For suggested schema of sensors installation See Fig. 2.

Based on accumulated experience of railroad catenary services, we may look ahead on further development of permanent monitoring systems:

1. Sensors advanced development with subsystems of electrical current leakage monitoring
2. Design of sensor for control of pantograph stress on catenary kit (digital video camera of high resolution would be perfect solution for visual assessment of actual pantograph condition).
3. Installation of temperature sensors per spots of feeding cables junction and on impedance bond with secondary winding plus dividers;
4. Sensors of partial discharge may be applied for isolation control within high voltage line of 10kW, (pair of cable – rail line), per area of alternating current (AC) (for the case of AC districts, it is critical to complete testing procedure for determination of sparks influence of sensors performance).



AS – anchorage section; OL – overhead line; CL – catenary line; CA – cables amplifier; VS – vibration sensor; SS – stressing sensor; L_{AS} – length; L – anchor length

Fig. 1. Schema of monitoring system sensors installation.

5. Arrangement of monitoring system for rigid and flexible beams in accordance with vibration features (angles of inclination per rigid beams with mechanical stressing for flexible beams).

6. Current leakage diagnostic must be ensured concerning discharge switches.

7. Current supervision shall be arranged per driveway cable, to assess the features of current spreading factor (it may be upgraded via additional control of connecting strings condition).

8. To develop the monitoring system for air railroad switches regarding vibration features, stressing, sparks formation etc.

9. Design the monitoring system for current collecting devices based on photo and video data analysis.

10. For the areas of direct current functioning, the system of insulation supervision shall be the matter.

11. Regarding sectional strain insulators monitoring system based on spark formation concert must be designed.

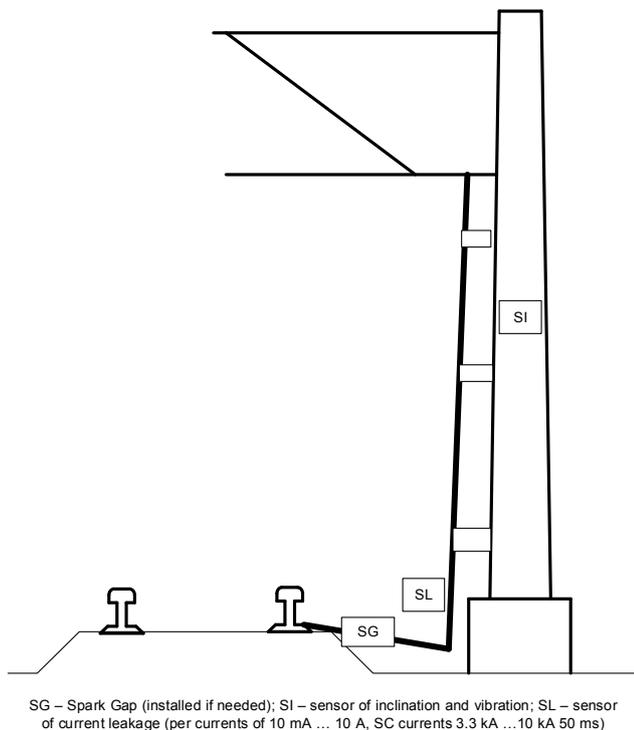


Fig. 2. Schema of sensors installation for masts condition monitoring.

Design of monitoring performance for power worked lever dividers in accordance with voltage and currents supervision.

III. BUSINESS-MODEL OF MONITORING SYSTEM PROMOTION

Substantiation of permanent monitoring system usually requires entrenchment of business-model with service crew quantity optimization nowadays. The idea of the above business-model is as follows: there is a set of schedule jobs to be completed with periodic sequence by servicemen team and in the event of monitoring system implementation, part of essential works considered to be automated, which allows us to optimize service brigade size.

Anyhow, during maintenance work performance via service crew with partial automation, accidents occurrence are being continued. Consequences of the aforesaid may be different, for example, serious failures; missed initial defects

while inspections or natural calamities. Unfortunately, even the entire computerization is not the panacea per failures occurrence at the monitoring site, but it can reduce those undesirable events of technological procedure violations.

The abovementioned business-model may not be the correct one in the event of substantiation. During automation performance implementation for the idea of man-factor reduction, enhancement of reliability is the function of expenses boost. For instance, introduction of microprocessor systems for railroad traffic management helps us to fulfill the itinerary arrangement automation (light signals with switches performance). Nevertheless, investment costs into the system with afterward periodical upholding are very high. For practical management point of view it is much more cost effective to place a pair of railroad switches per station with couple of operators, who shall cope those switches manually. In this particular case, automation factor, anyway, shall be expensive one, which looked-for investment into service with maintenance performance, but some of those savings shall be paid off later.

While railroad permanent monitoring systems presentation we may apply the other business-model, for example, Designer of the system may sell not the technical equipment itself, but the 'event factor', such as the 'pre-failure status quo' (either 'accident prevention' or 'technical function prevention' etc.). In this case, the Designer may apply his proper technical instruments to achieve the ordered task, and the Client shall receive the needed commodity such as safe and non-stop technological performance! Per the case of railroad catenary such a business-model looks like the attractive one, because for the danger of the malfunction incident, the outcome must be long term gaps within train schedule together with infrastructure damages plus rolling stock breakage. The benefit of our suggested modeling arrangement approach is that the Designer may take the total responsibility of the entire life cycle of the whole monitoring system at site.

In our business-model, expenses of the monitoring site owner per the initial stage is the matter of the minimization for the reason, that monitoring kit for him is the instrument of information accumulation with safety improvement only. While service time continuation period, when failures may appear, the effect shall be considered feasible for the Designer.

IV. CONCLUSION

Methods recommended by authors regarding functional features options broaden of permanent monitoring systems allows us in fact to cover the whole spectrum of diagnostic features with appeared technological situations within railroad catenary. Actual realization of those described options helps us to upgrade drastically the outcome level of permanent monitoring system per railroad catenary. We should add, that the set of focused ways of permanent monitoring of railroad auxiliary elements evolution can be applied for the purposes of urban electrical transportation monitoring (for sure, including available specifics).

Diagnostic devices installation methods proposed by the authors allow obtaining the necessary diagnostic data range sufficient to conduct high-quality monitoring of the railway contact-wire of catenary system parameters. The monitoring results at the first stage are taken into account in the organization and planning of maintenance and repair work in the electrification and power supply department of railways. At

the next stages, the monitoring results must be taken into account by train dispatchers and train driver to select the most rational ways of traffic control, taking into account the occurrence of emergencies with a contact-wire of catenary system.

In conclusion, we would like to mention the future digital approach to the net railroad monitoring systems [19 – 22]. As well we do believe in the future feedback of rolling stock monitoring automation shall run, not to dispatcher only, but to train operators to ensure professional situation assessment to avoid off-nominal conditions.

REFERENCES

- [1] Z. Liu “Detection and Estimation Research of High-Speed Railway Catenary”, Springer Nature Singapore Pte Ltd, 2017, 287 p.
- [2] D.V. Barch “Improvement of the Service System Based on Monitoring and Diagnostics of Power Supply Equipment”, Proc. of Petersburg transport university, 2012, Issue 3, pp. 103-110. (in Russian).
- [3] “Dynamic Catenary Monitoring DCM”, Furrer+Frey, 2012, 12 p.
- [4] J. Yu, and M. Wu “Development of a Detection System for the Catenary Vibration Monitoring”, International Conference of Information Technology, Computer Engineering and Management Sciences, 2011, Vol. 1, 24-25 September 2011, Nanjing, Jiangsu, China, pp. 76-79, doi: 10.1109/ICM.2011.155.
- [5] H. Hofler, M. Dambacher, N. Dimopoulos, and V. Jetter “Monitoring and Inspecting Overhead Wires and Supporting Structures”, IEEE Intelligent Vehicles Symposium, 14-17 June 2004, Parma, Italy, pp. 512-517, doi: 10.1109/IVS.2004.1336436.
- [6] N. Theune, T. Bosselmann, J. Kaiser, M. Willsch, H. Hertsch, and R. Puschmann “Online Catenary Temperature Monitoring at New High-Speed Rail Line Cologne-Rhine/Main”, WCRR, 2003, Vol. 18, Issue 5, pp. 1038-1043.
- [7] Y. Park, Y.H. Cho, K. Lee, H. Jung, H. Kim, S. Kwon, and H. Park “Development of an FPGA-based Online Condition Monitoring System for Railway Catenary Application”, 8th World Congress on Railway Research, COEX, Seoul, Korea, 2008, 18-22 May.
- [8] Y. Park, S.Y. Kwon, and J.M. Kim “Reliability Analysis of Arcing Measurement System Between Pantograph and Contact Wire”, The Transactions of the Korean Institute of Electrical Engineers, 2012, Vol. 61, No. 8, pp. 1216-1220.
- [9] T. Hisa, M. Kanaya, V. Sakai, and K. Hamaoka “Rail and Contact Line Inspection Technology for Safe and Reliable Railway Traffic”, Hitachi Review, 2012, Vol. 61, Issue 7, pp. 325-330.
- [10] M. Mizan, K. Karwowski, and D. Karkosiński “Monitoring odbieraków prądu w warunkach eksploatacyjnych na linii kolejowej”, Przegląd Elektrotechniczny, 2013, R89, nr. 12, pp. 154-160.
- [11] Y. Park, K. Lee, C. Park, J.-K. Kim, A. Jeon, S. Kwon, and Y.H. Cho “Video Image Analysis in Accordance with Power Density of Arcing for Current Collection System in Electric Railway”, The Transactions of the Korean Institute of Electrical Engineers, 2013, Vol. 62, Issue 9, pp. 1343-1347.
- [12] “Sicat CMS. Catenary Monitoring System for Overhead Contact Line Systems”, Product information, Version 1.1.4, Siemens AG, 2016, 8 p.
- [13] H. Wang, A. Núñez, Z. Liu, J. Chen, and R. Dollevoet “Intelligent Condition Monitoring of Railway Catenary Systems: A Bayesian Network Approach”, The 25th International Symposium on Dynamics of Vehicles on Roads and Tracks, 14-18 August 2017, Rockhampton, Australia, pp. 1-6.
- [14] D. Efanov, G. Osadchy, D. Sedykh, D. Pristensky, and D. Barch “Monitoring System of Vibration Impacts on the Structure of Overhead Catenary of High-Speed Railway Lines”, Proc. of 14th IEEE East-West Design & Test Symposium, Yerevan, Armenia, October 14-17, 2016, pp. 201-208, doi: 10.1109/EWDTS.2016.7807691.
- [15] D. Efanov, G. Osadchy, and D. Sedykh “Development of Rail Roads Permanent Monitoring Technology Regarding Stressing of Contact-Wire Catenary System”, Proc. of 2nd International Conference on Industrial Engineering, Applications and Manufacturing, Chelyabinsk, Russia, 19-20 May, 2016, pp. 1-5, doi: 10.1109/ICIEAM.2016.7911431.
- [16] D. Efanov, D. Sedykh, G. Osadchy, and D. Barch “Permanent Monitoring of Railway Overhead Catenary Poles Inclination”, Proc. of 15th IEEE East-West Design & Test Symposium, Novi Sad, Serbia, September 29 – October 2, 2017, pp. 163-167, doi: 10.1109/EWDTS.2017.8110142.
- [17] A. Belyi, G. Osadchy, and K. Dolinskyi “Practical Recommendations for Controlling of Angular Displacements of High-Rise and Large Span Elements of Civil Structures”, Proc. of 16th IEEE East-West Design & Test Symposium, Kazan, Russia, September 14-17, 2018, pp. 176-183, doi: 10.1109/EWDTS.2018.8524743.
- [18] A.A. Kovalev, V.M. Masloc, and N.A. Aksenov “The Use of Mobile Devices for Catenary Supports Diagnostics”, Transport of the Ural, 2018, Issue 2, pp. 77-79 (in Russian).
- [19] “Digital Railway Strategy”, Network Rail, April 2018, 42 p.
- [20] T. Bauer, and D.N. Benito “Digital Railway Stations for Increased Throughput and a Better Passenger Experience”, Signal+Draht, 2018, issue 7+8, pp. 6-12.
- [21] R. Stäuble, and P. Gschwend “Digital Signalling in the Simmental”, Signal+Draht, 2018, issue 10, pp. 40-46.
- [22] D. Efanov, and G. Osadchy “Paradigms for Building Control Systems on Railroad Transport: from the Systems of Electrical Interlocking of Points and Light Signals to Smart Grid Train Movements Controlling Systems”, Proc. of 16th IEEE East-West Design & Test Symposium, Kazan, Russia, September 14-17, 2018, pp. 213-220, doi: 10.1109/EWDTS.2018.8524809.

Design of real-time system logic control on FPGA

Maryna Miroshnyk
dept. Specialized Computer Systems
Ukrainian State University of Railway
Transport
Kharkiv, Ukraine
marinagmiro@gmail.com
0000-0002-2231-2529

Olexander Shkil
dept. Computer Engineering Design
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
oleksandr.shkil@nure.ua
0000-0003-1071-3445

Elvira Kulak
dept. Computer Engineering Design
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
elvira.kulak@nure.ua
0000-0002-8441-5187

Dariia Rakhlis
dept. Computer Engineering Design
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
dariia.rakhlis@nure.ua
0000-0002-6652-1840

Inna Filippenko
dept. Computer Engineering Design
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
inna.filippenko@nure.ua
0000-0002-3584-2107

Maksym Hoha
dept. Computer Engineering Design
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
maksym.hoha@nure.ua

Mykyta Malakhov
dept. Computer Engineering Design
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
mykyta.malakhov@nure.ua

Vladyslav Sergienko
dept. Computer Engineering Design
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
vladyslav.serhiienko@nure.ua

Abstract— Problems of real-time hardware logic control systems design on the FPGA are considered. The control algorithm is implemented based on a timed FSM model, represented by a temporal state diagram. The design of the control device model using hardware description language VHDL in the form of the three-process pattern is made. The functional verification of the model was carried out using Active-HDL tools, the synthesis of the circuit was carried out on the Spartan 3E FPGA technology platform using Xilinx ISE CAD tools. The hardware costs for the circuit implementation of the control device were analyzed.

Keywords— timed FSM, temporal state diagram, VHDL, functional verification, pattern, FPGA.

I. INTRODUCTION

Among the entire set of control systems, the significant part are logical control systems, in which control signals take values of the logical zero or one, depending on boundary values of physical quantities that define these parameters. For the technical implementation of these systems, the Finite State Machine (FSM) is the most suitable, and the visual representation of functioning algorithm is a state diagram. The distinctive feature of the FSM for logic control is that among input values there are not only announcing signals of the operational state machine, but also external, towards to the controlled system, events of external world, which are playing role of interrupts for the control algorithm.

A control FSM functions in machine time, is determined by the operation time of machine. But the most of real logical control systems cooperate with external world in the metric time, i.e. they are real-time systems.

The real-time control system is a system in which the resultant action (activity) depends not only on logical values of simple control actions, but also on time during which these actions are performed. The main difference between tasks in real time and tasks that are not dependent on time is that tasks in real-time systems must be completed within a specified

period of time, that allow to complete processing of data correctly. For their implementation, it is customary to use a timed FSM model, which allows taking into account the effect of metric time on transitions between technical states of control system.

Any local digital device that implements an information processing or control algorithm can be implemented in two ways: hardware or software-hardware. With hardware implementation method, a given algorithm is described in hardware description language (HDL) and is synthesized by instrumental tools of computer-aided design (CAD) in FPGA (Field-Programmable Gate Array circuit). The advantage of this approach is hardware flexibility (ability to implement any algorithm) and a sufficiently large speed.

During describing the functioning algorithm for digital logical control devices in CAD systems, one of code styles is the style of automata-based programming. In automata-based programming, a concept of "state" is used as the base one [1]. A state is a mathematical abstraction that is uniquely associated to each of physical states of a control object, since usually an operation of technical systems is shown through a change of their states. At the same time, each state in a control algorithm maintains a control object in a proper state, and the transition to a new state in an algorithm leads to the transition of an object to a new corresponding state, which ensures the process of object' logical control. A state is a set of parameters of a technical system at a given moment of time. A current state carries all information about the history of a system, which is necessary to determine its response to any input action that is formed at a given time.

Thus, the task of developing an unified pattern in the hardware description language for the design of real-time logic control devices, which based on FSM in the style of automata-based programming, becomes urgent. The goal of this work is to develop a pattern for describing finite state

machine in the hardware description language VHDL, and automated synthesis of the received model with.

II. THE MODEL OF STRUCTURAL FSM IN REAL-TIME SYSTEMS

When describing a behavior of real-time control systems, it is necessary to take into account timing aspects of their behavior. For this, a state machine model is expanded by introducing a timed variable, and the concept of a timed FSM [2, 3] is introduced. A timed variable constantly increases its value and "resets" to 0 upon the arrival of an input signal and a FSM transitions to a new state. Time variables are measured in automata cycles.

As a rule, three parameters are used to describe timing aspects in the automata-based model: timing constraints t_c , (input) timeouts t_{io} and output delays t_d , which sometimes are called as output timeouts. An input timeout determines the maximum waiting time for input effects (events) for each state of a FSM. If an input symbol was not filed before the end of a timeout, a state machine starts polling input variables and can switch to another state. Time constraints are intervals on transitions that limit the time during which the transition can be performed. Output delays (output timeouts) shows the time that a state machine spends on executing of a transition, i.e. an output signal will appear at an output after a time interval, which is determined by the output delay.

In logical control systems, a concept of "input values" is divided into input actions and events. Input actions are implemented automatically by polling in accordance with an algorithm of its operation in a control loop, and events are implemented instantaneously and lead to a change in a state of the state machine.

Event processing in real-time systems, as a rule, are determined on a basis of dynamic characteristics of control processes and related events. An event is an abstract concept, implying such a change in environmental conditions, which generates a certain reaction of a system [4]. Events can be generated both by an external environment and within a control system by its components.

There are three main options for an interaction of a control FSM with an external environment.

1. Events are used for an interaction of a control and operating FSM within an automatic control system. In this case, if events are exceptional (two events cannot occur simultaneously), events' processing doesn't differ from processing of input variables values of a FSM.

2. Events along with input variables provide an interaction of a FSM with an external environment. This design solution should reflect the difference between events and input variables: a FSM processes events at the moment of its occurrence, while values of input variables are polled by a FSM on its own initiative.

3. An each event is associated with a separate state (transition) of a FSM. This solution is only suitable for implementing of an exceptional event model. In addition, it reflects an active role of events, and the fact that the occurrence of events, by itself, initiates an operation of a FSM. This solution is the best coordinated with traditional event systems, where any output function is related to the content of events.

Depending on a purpose and features of using models of a timed FSM, there are many modifications of such models, which differently take into account both, the method of events' processing and the way of delays' accounting in states of a FSM [5, 6].

Based on functioning features of logic control systems, a full model of a structural timed FSM can be represented by a nine $W = (X, Y, Z, f, g, z_0, T_c, T_{io}, T_d)$, where: $X = \{X_C, X_E\}$ – a set of input variables, X_C – a set of announcing signals from a control object, X_E – a set of external events; $Y = \{Y_C, Y_F\}$ – a set of output variables, Y_C – a set of reactions (control signals), Y_F – a set of activities (output functions); Z – a set of internal variables that determine coding states of a FSM; f – a transition function, g – an output function; z_0 – a code of the initial state of a FSM; $T_c = \{t_{c1}, t_{c2} \dots t_{cp}\}$ – a set of timed variables for timing restrictions on each arc of a state diagram, where p – is a number of arcs in a state diagram, $t_{ci} = \{1, k\}$, k – a maximum number of clocks' restrictions on transitions to the i -th node of a state diagram in polling mode, $k = \{1, \infty\}$, ∞ – responds exclusively by an effective transition function, $T_{io} = \{t_{io1}, t_{io2} \dots t_{ion}\}$ – a set of timed variables for timeouts (expected) of each state of a FSM, $t_{toi} = \{1, n\}$ – timeout for each state, n – a number of states of a FSM; $T_d = \{t_{d1}, t_{d2} \dots t_{dm}\}$ – is a set of delays for the realization of the corresponding output signal, where m – is a number of output variables, $t_{dm} = \{1, l\}$, where l – is a maximum number of clock cycles for the realization of output functions in the specified state of a FSM.

In general, a timed FSM can contain all three time parameters, but for a specific task timed FSM with one or two of specified parameters can be used.

A classical model of timed FSM, which consist of three timing parameters $\langle t_c, t_{io}, t_d \rangle$ can't be directly attributed to the traditional Moore model. The output function is similar to Moore FSM, but the output signal is formed after delay, and not when the FSM transits to a new state. A time of appearance (change) of output signals is connected to a working edge of the synchronization signal. In the proposed model of the timed FSM, the logic of its operation is as follows.

During FSM transitions to the current state a_i , the main time parameter $t_{io}(a_i)$ (timeout) is determined for it, that is, a time during which a FSM should be in the current state if an external event will not transfer the FSM into another state ahead of time. Value of t_{io} is defined in FSM cycles. After the time t_{io} is expired, a FSM responds to input signals (polls them) and transfers to a next state. Output signals of a FSM in the current state a_i appear at outputs of the FSM at the time determined by $t_{dj}(a_i)$ (output delays), that is, output delays for signals y_j in the state a_i . For each of output signals y_j , the initial delay is determined in FSM cycles and can be different. When $y_j = 0$, timed FSM approaches the classical Moore model.

A processing of external events is as follows. For each state a_i , the time constraints $t_c(a_i)$ (input constraints) are set,

that is, a time interval during which a FSM, staying in the state a_i , can process initial events. Timing constraints are determined in FSM clock cycles and calculated as $t_c = (t_1 - t_0)$, where t_0 – is the beginning of timing constraints’ “window”, t_1 – is the end of timing constraints’ “window”. When $t_1 = \infty$, a timed FSM without input timing constraints is considered. If an external event occurs outside of the "window" of timing constraints, a state machine does not respond to it.

Logic control devices, based on FSM, function in a FSM time, which is measured in FSM clock cycles, i.e. discrete time intervals during which a FSM transfer from one state to another. The duration of a FSM cycle in real devices is usually determined by the frequency of the clock signal Clk . A temporal state diagram is used to describe a timed FSM. All timing parameters of a temporal state diagram are implemented through loops. Conditions for those loops are counting of the number of clock cycles Clk , which is implemented by the counter (*count*) in the FPGA [7]. The fragment of the temporal state diagram for three states is shown in fig. 1.

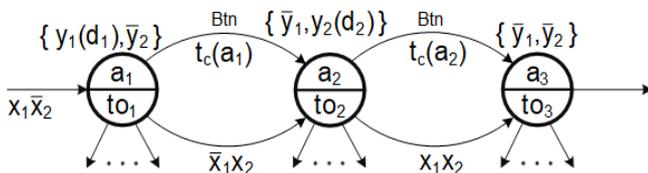


Fig. 1. Fragment of the temporal state diagram of Moore FSM

Figure 2 shows the fragment of the timing diagram of a FSM functioning.

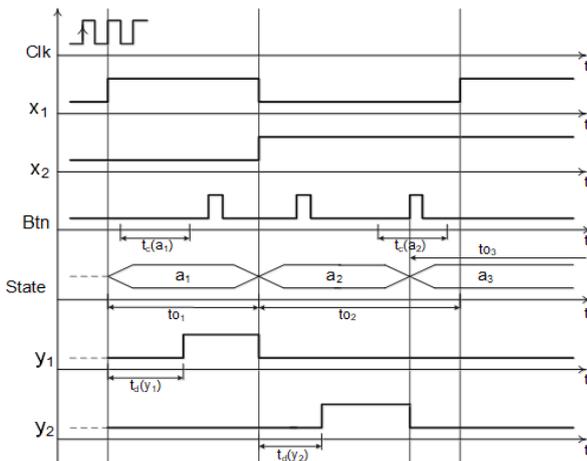


Fig. 2. The timing diagram of a timed FSM

III. AN EXAMPLE OF AN AUTOMATED DESIGN OF TIMED FSM ON FPGA

When designing an operation algorithm of a digital device in HDL, it is important that the developed HDL code doesn't go beyond limits of the synthesized subset of the particular HDL. Single-process and two-process patterns for design of HDL-models of Moore timed FSM with delays in states are considered in [7, 8].

On the one hand, a single-process pattern is correctly synthesized for Moore FSM, but it generates hardware redundancy for Mealy FSM (register for output signals is synthesized). On the other hand, a two-process pattern, taking

into account the counter signal that implements the delay, is not synthesized correctly. Therefore, for the implementation of a VHDL model of a timed FSM, it was proposed to use a three-process pattern: a synchronous process of new state assigning, a synchronous process of implementing FSM clock' counter and a combinational process of transition function implementing. Outputs function of Moore timed FSM is implemented through the conditional signal assignment statement out of processes.

As an example of the implementation of proposed structure of a timed FSM' HDL-model, let's consider the temporal state diagram of the modified Moore FSM, which is represented in fig. 3. In this state diagram, x_1 and x_2 are considered as input actions, and Btn is considered as an event.

Timing parameters for the temporal state diagram, which is preset in FSM cycles, are as follows:

- input constraints for Btn : [3; 3] for a_1 , [4; 5] for a_2 ;
- timeouts for states: $T_1=6, T_2=7, T_3=9, T_4=5$;
- output delays for signals: $y_1(d_1)=1, y_2(d_2)=2$.

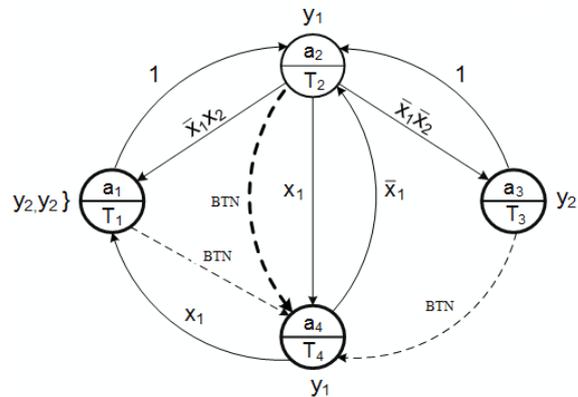


Fig. 3. State diagram of Moore FSM for control device

In the considered state diagram, the signal Btn – is an event that is, essentially, the input signal with the highest priority. In this regard, in the pattern during description of transitions from considered state it is checked firstly in the IF branch. In all other transitions from this state (elsif ... else branch) this signal is equal \overline{Btn} and is not explicitly written. Therefore, for greater clarity, the arc with Btn is highlighted by a dotted line on the state diagram, and \overline{Btn} is not present in the expressions of the transition conditions, that is, the orthogonalization of transition conditions is not violated here.

Figure 4 presents fragments of the VHDL model corresponding to the temporal state diagram in Figure 3. Here state synchronization – is the process of new state assigning, timer synchronization – is the process of the FSM clock' counter implementation, transition function – is the combinational process of the transition function implementation, output function – is the conditional assignment statement for output signals.

-- state synchronization

```

process (Clk, Reset)
begin
if Reset = '1' then state <= a1;
elsif rising_edge(Clk) then state <= next_state;
end if;
end process;

```

```

-- clk synchronization
process (Clk, Reset)
begin
if Reset = '1' then count <= (others => '0');
elsif rising_edge(Clk) then
if State /= next_state then count <= (others => '0');
else count <= count + 1;
end if;
end if;
end process;
-- transition function
process (state, x, Btn, count)
begin
case State is
...
when a2 =>
if Btn = '1' and count >= constraint_a2_L - 1 and
count < constraint_a2_H then next_state <= a4;
elsif count < T2 - 1 then next_state <= state;
elsif x(1) = '1' then next_state <= a4;
elsif x(2) = '1' then next_state <= a1;
else next_state <= a3;
end if;
-- output function
y( 1 ) <= '1' when ( ( state = a1 ) or ( state = a2 ) or ( state =
a4 ) ) and count >= output_delay_Y1 else '0';

```

Fig. 4. Fragment of the VHDL model of timed Moore FSM

Figure 5 shows the timing diagram (waveform) of the simulation results of the considered control device of the ALDEC Active-HDL system.

The processing of the *Btn* event at time 2550 ns is of particular interest. This event falls into the interval *Input Constraint* for *Btn* [4; 5] for the state *a2* and realizes the transition to the state *a4* (highlighted arc *Btn* in fig. 3).

The synthesis report in figure 6 shows the results of the synthesis of the control device in the XILINX ISE system for FPGA: Spartan 3E, XC3S500E chip, Package FG 320 (xc3s500e-4fg320).

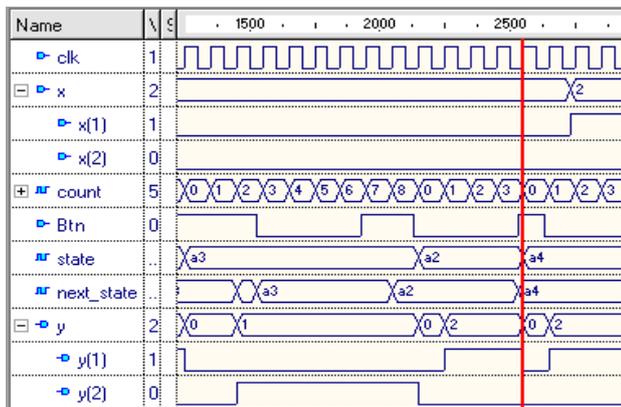


Fig. 5. Control device simulation results

```

Analyzing FSM <MFsm> for best encoding.
Optimizing FSM <FSM_0> on signal <state[1:2]> with user
encoding.
State | Encoding a1 | 00 a2 | 01 a3 | 10 a4 | 11
Final Register Report : Macro Statistics
# Registers : 6 Flip-Flops : 6

```

Fig. 6. The synthesis report of control device

The structure, consisting of two blocks, is synthesized: control FSM (2 D flip-flops for states coding, combinational

circuits implementing transition and output functions) and counter based on 4 D flip-flops for counting 9 cycles of the maximum timeout. To confirm the complete correctness of timing parameters of the proposed model, it was necessary to perform timing simulation, but this is the subject of further research.

CONCLUSION

As a result of the conducted research, it was shown that during automated design of real-time logic control systems it is advisable to use models of the timed control FSM. Problems of constructing timed FSM that take into account timing constraints, input timeouts and output delays were considered. To describe these models in hardware description language VHDL during automated design, a three-process pattern in the style of automata-based programming for Moore FSM was developed, which contains the combinational process for describing transition functions, the synchronous process for new state assigning, and the synchronous process for accounting of FSM cycles. The simulation of developed VHDL model in the Active-HDL system and the circuit synthesis using the XILINX ISE CAD tools in the FPGA on the Spartan 3E board showed the efficiency of the proposed model. At the same time, hardware costs don't go beyond the standard rate for FSM states' encoding and formation discharges of FSM cycles' counter.

A practical value of obtained results is that authors proposed the pattern, describing algorithms for the functioning of the timed FSM in real-time logic control systems in the VHDL language, which can be used by beginner designers of digital logic control systems, as well as students of the specialty "Computer Engineering".

A direction of further research may be the use of the Mealy model for the implementation of timed control FSM

REFERENCES

- [1] Shalyto A. A. Logical control. Hardware and software implementation methods/ A. A. Shalyto. – StPet.: Science, 2000. – 780 p.
- [2] Alur R. A theory of timed automata / R. Alur, D.L. Dill // Theoretical Computer Science. – 1994. – V.126. – N 2. – P. 183-235.
- [3] Merayo M.G. Formal Testing from Timed Finite State Machines / M.G. Merayo, M. Nunez, I. Rodriguez // Computer Networks. – 2008. – Vol. 52– №2. – P. 432-460.
- [4] Matushin A.O. Programming microcontrollers: Strategy and tactics / A. Matushin. – M.: DMK Press, 2017. – 356 c.
- [5] Zhigulin M. FSM-Based Test Derivation Strategies for Systems with Time-Outs / M. Zhigulin, N. Yevtushenko, S. Maag, A.R. Cavalli // QSIQ 2011. – P. 141-149.
- [6] Gromov M. Testing Components of Interacting Timed Finite State Machines / M. Gromov, A. Tvardovskii, N. Yevtushenko // Proceedings of IEEE East-West Design & Test Symposium (EWDTS), 2016. – P. 193–196.
- [7] Shkil A. Design of Logical Control Units Based on Finite State Machines' Patterns/ M Miroshnyk; S. Poroshyn; A. Shkil; E. Kulak; I. Filippenko; D. Kucherenko; Y. Pa khomov; J. Salfetnikova; M. Goga//Proceedings of the 2018 IEEE East-West Design & Test Symposium.– 6 p. [Електронний ресурс] / IEEE Xplore Digital Library. – Режим доступу: www / URL: <https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?filter=issueId%20EQ%20%228524135%22&refinements=Author:Alexander%20Shkil&pageNumber=1&resultAction=REFINE>.
- [8] Shkil A. /Design automation of testable finite state machines. / Miroshnik M.A., Shkil A.S., Kulak E.N., Filippenko I.V., Kucherenko D.Y., E.E. German // Proceedings of IEEE East-West Design & Test Symposium (EWDTS), 2017. – P. 203-208. .

Emerging Culture of Social Computing

Anastasia Hahanova
Design Automation Department
Kharkov National University of
Radioelectronics
Kharkov, Ukraine
hahanova@icloud.com

Svetlana Chumachenko
Design Automation Department
Kharkov National University of
Radioelectronics
Kharkov, Ukraine
svetachumachenko@icloud.com

Vladimir Hahanov
Design Automation Department
Kharkov National University of
Radioelectronics
Kharkov, Ukraine
hahanov@icloud.com

Abdullayev Vugar Hacimahmud
Computer Engineering
Department
Azerbaijan State Oil and Industry
University
Baku, Azerbaijan
abdulvugar@mail.ru

Ka Lok Man
Xi'an Jiaotong-Liverpool University
China
kalok2006@gmail.com

Alexander Mishchenko
Design Automation Department
Kharkov National University of Radioelectronics
Kharkov, Ukraine,
USA
santific@gmail.com

Abstract— Social computing is proposed to improve the quality of life of citizens and preserve the ecology of the planet through moral digital and human-free management of each citizen based on accurate monitoring of their preferences. The model of cyber social computing is described, which will save humanity from its vices and direct the efforts and minds of people to moral solving the existing problems of energy, materials, ecology, quality of life. We propose solutions to the problems of managing social groups from the perspective of deterministic metric computing instead of management based on statistical analysis that does not take into account the interests of each individual citizen. Democracy is an anti-scientific method of probabilistic management of social groups and making incompetent decisions. It should be replaced by moral cyber-computing of metric management of society based on exhaustive testing of the interests of every citizen. The strategy of uniting social groups is shown to effectively address economic and industrial problems based on the adoption of a constitution and laws, which integrate citizens' efforts through the implementation of the doctrine of the unity of diversity of languages, religions, histories, traditions and cultures.

Keywords— *emerging culture, social computing, human-free management, social process management, metric of relationships*

I. INTRODUCTION

As cyberspace develops, the physical world transforms from dominant to subordinate. All physical processes and phenomena today have their own digital images, which are gradually transformed into prototypes, and the real world is becoming increasingly vulnerable, dependent and controlled from the virtual cyber world. The following axiom should be considered: who leads in cyberspace rules the physical world. The moral fact is that the cyber-physical world positively connects all the inhabitants of the planet with each other without intermediaries through social networks, cloud services and Edge Computing. The state has lost its monopoly on the delivery and distribution of information among its citizens. The censorship of information flows in cyberspace from degraded state institutions, does not have even the slightest right to exist. Giving public information or hiding it means that the author has a desire to receive certain moral or material dividends.

The IEEE Xplore library has practically no publications in the field of Cyber Social Business Computing, and Springer has 13358 books. At the same time, IEEE Social Computing includes 25342 works, and Springer is represented by 41,733 monographs. Naturally, the combination of two market-focused research topics can provide a significant practical result in improving the quality of business, life and preserving the ecology of the planet. There is only one Springer book [1], as a collection of articles on the results of the same scientific seminar, indirectly affecting the management of the cyber-physical world. Characteristic publications [2-4] are focused on social networks and monitoring citizens' preferences without generating control actions automatically. Therefore, the article [5], monograph [6], their development and improvement, devoted to active cyber-physical computing related to the active management of social, business processes and phenomena based on their accurate monitoring is very timely and relevant. The market of services in cyberspace is still using the "cave wall" information display systems designed for the eyes of a person, who is given the function of making, as a rule, erroneous actuative decisions leading to social conflicts, economic and financial losses. Disposing a person from the function of managing a socially-oriented business and transferring it to cyber-physical human-free business computing is the most important organizational problem of the moral creative world. A person is not able to accurately control even himself, constantly forgetting his historical experience, he regularly steps on the same rake of past mistakes. Therefore, a citizen, social group, company, state and humanity need to create a scalable Gartner-computing avatar: "virtual assistant - digital twin - smart robot", which will save people from wrong decisions leading to undesirable consequences in the business and market of social technologies.

The goal of the research is to create structural metric cyberculture of computing, as a technology for deterministic digital control of cyber-physical and social processes and phenomena based on accurate and comprehensive monitoring of the state of cyberspace.

Research objectives: 1) Definitions and axioms of computing. 2) The relation is the fundamental principle of cyber-physical and social processes and phenomena. 3) Democracy is the ignorant computing for managing social processes. 4) Cyber social computing for solving problems of social management.

II. DEFINITIONS AND AXIOMS OF COMPUTING

Computing is a branch of knowledge that develops the theory and practice of reliable metric management of virtual, physical (natural) and social processes and phenomena based on the use of data centers of big data through digital monitoring of cyber-physical space using intelligent services for retrieval and analysis, personal gadgets and smart sensors.

Systematically, computing (Fig. 1) is the process of monitoring (5) and activating (6) metric relations (2) in the infrastructure of management (3) and execution (4) to achieve and visualize (8) a goal – product (1) with specified resources (7).

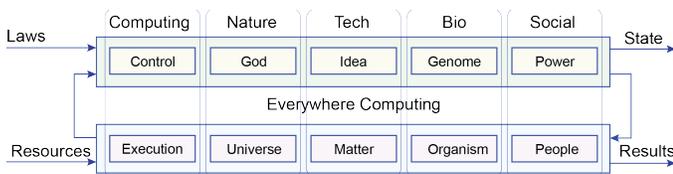


Fig. 1. Computing

The metrical and structural definition of computing through eight interconnected components provides a theoretical fundamental basis for the formal and actual creation of any process in a given field of human or natural activity. Types of computing on the entered metric cover all fields of human activity: cosmological, biological, floristic, physical, virtual, quantum, social, state, medical, transport, infrastructure, scientific, educational, industrial, sports, recreation, travel, entertainment.

The process is material and energy interaction of system components in time and space to achieve the goal. Globally, the process is a material and energy change in the space-time continuum. Locally, the process is the development of the spatial relationship of components (phenomena) in time.

A phenomenon is a component (system) or a fragment of a process at a fixed moment or interval of time perceived by sensors, feelings, faith, or mind.

The computer is a process, the observed interaction of control and execution mechanisms in time and space based on monitoring and actuation of metric relations to achieve a goal in the form of products or services for given resources.

III. RELATION IS THE FIRST PRINCIPLE OF PROCESSES AND PHENOMENA

Relation is a structure of interconnected components that determines the properties of a process or phenomenon. The structure determines the properties of the components, process or phenomenon, but not vice versa. Relation-signature is primary, media components are secondary. The alphabet is the carrier of a relation defined by operations (signatures) on

symbols. The symbols of the alphabet alone do not make sense. Relations are decisive in creating effective mathematical theories, data structures, algorithms, architectures, models, methods, technologies, materials, services, software and hardware applications, cyber-physical and social systems, including economics, healthcare, transportation, law and order, ecology and statehood. The cardinality of the relation as an integral set and the quality of mutual relationships between the components forms a metric that makes it possible to identify the effectiveness of the structure.

“In the beginning was the Word, and the Word was with God, and the Word was God” (the Gospel of John). The word as a relationship between people, processes or phenomena determines their properties, but not vice versa. In the system, the relation-word is primary, the carriers are secondary: people, components. Relationships form the efficiency and viability of the system, company, state. The idea of the development of a process or phenomenon is primary, its material implementation is secondary. The Universe develops according to the idea conceived by God (Nature), realized by the relations in pairs: matter-energy, space-time. The Creator (Laws, Idea) is primary, the Universe is secondary. The scaling of the axiom of the primacy of an idea in the material world leads to a new understanding of the following facts: relations and social laws are primary, their implementation in society is secondary. A partial change of the above words from the Gospel may be represented in a modern scientific style: "In the beginning was the Idea, and the Idea was with God, and the Idea was God." The idea is a formulated algorithm, program or genome of system development, as a set of relations. The idea always comes from the Creator, who brings it to the material world. The idea is God who concentrates all the laws of the material-energy and space-time development of the Universe. The laws of the Universe or God exist independently of human. The goal of Homo Sapiens is to know God, Nature; it means to discover the laws existing independently of human and to morally use them for the benefit of humanity, preserving Nature.

“In the beginning was the Word” means that the Word or the genome of the development of the Universe is the beginning and does not depend on the level or phase of the development of the Universe, the living Nature and Humanity. “The word was with God” means that a set of Laws-Ideas for the development of the Universe is concentrated under the name of the Creator or God. “The word was God” means that naturally all the Laws or Relations of the development of the Universe, humanity, and living Nature, exist independently of us and are the First Cause or God. Relations, Laws, God are invariant (non-material) to matter, the carrier of relations or the components involved in them are material. Teleportation of a person into the Universe is also covered with the thesis “In the beginning was the Word”. In order to expand life in outer space, it is necessary to find a suitable environment and transfer the Word to it as a genome or algorithm for the development of living matter.

Materialism is based on the doctrine of the primacy of matter in relation to an idea. Given the primacy of the relationship that works on the pair: “idea – matter,” such a statement cannot be perceived as truth. The main role in this process-computing relation belongs to the idea that programs or predicts the development or creation of the material world. How then to be

with a pair of "chicken-egg"? Here, the root cause is the genome (law) of the development of relations between them, and the material implementation of the genome in the form of chicken and egg phenomena is secondary.

Relationship symmetry, as two wings, plays a positive role in the development of the Universe, humanity and an outstanding Person. If you are a genius or talent and you do not see your counterpart (envious, opponent), then look for or create him, because of without him you will not take place! When you are criticized, you have many enemies, then you are worth something. "Consent and recognition rarely encourage moving forward and seeking" (Thor Heyerdahl).

IV. DEMOCRACY IS CLUELESS COMPUTING FOR SOCIAL PROCESS MANAGEMENT

The system essence of citizens in cyber-physical social computing is defined as the execution mechanism that creates services and/or products. At the same time, the management mechanism is formed by experts with relevant education and management experience in various fields of human activity. As a rule, the mentioned mechanisms should not intersect in human resources. Democracy puts the structure of computing, which is trivial to understand, upside down. The most ignorant in the management part of the population is involved to the management of the company, university, state and making strategic decisions, in fact, turning the role of professionally trained managers to zero. The classics of social computing are formed by the following components: 1) relationship-laws, 2) management, 3) execution, 4) infrastructure, 5) observation, 6) visualization, intended for manufacturing products and services subject to the availability of resources. Democracy is always an effective tool for making incompetent decisions by the ignorant political elite of the totalitarian state. The metric of democracy is its corruption, bribery and intimidation of citizens, the false justification for the seizure of power by a group of people who do not have moral goals and ways to achieve them.

Democracy, as a technology of incompetence, is the cause of all the ills of humanity: 1) A fig leaf on the bare body of managerial ignorance. 2) The greatest nonsense invented by humanity to justify almost always incompetent decisions. 3) Dice, where millions of lives are bet. 4) True democracy experts are in the minority and always lose. The essence of democracy is to make decisions by voting ignorant people in the government.

The historical roots of democracy are a cave decision-making technology, where everyone has the same universal knowledge about the issue of voting and its possible consequences. Today, society is a collection of deeply specialized citizens, each of whom is focused on the optimal solution of specific problems of a narrow profile. Therefore, the use of a random person or a group of incompetent persons to make a decision on the choice of a leader is nonsense, associated with the historical inertia of false thinking and the fake involvement of the masses in the government. The alternative to democracy is autocracy or the authoritarian power of the minority over the majority. With competent knowledge and morality, the head of an authoritarian regime of government is more effective for citizens of the state in terms of the quality of life and the preservation of the environment in the territory.

"Voting in science cannot solve problems, since there is always a majority of mediocrity, and there are only a few talents" (quoted from the film "I'm going to a thunderstorm"). The ignorance and inconsistency of democracy in science have ruined thousands of projects, casting aside the development of humanity tens of years ago. Only the authoritarian rule of experts, developing into cyber administration, is the key to the technological future of humanity.

Decision-making in a specialized board for awarding academic degrees is based on the voting of several experts in a particular field, which is a giant step in the development of a new social management technology based on digital monitoring of processes and phenomena. The metrical elections of the president of the country performed not by citizens, but experts in this field, also represent a more advanced technology in comparison with the democracy of incompetent citizens. However, the optimal model of cyber-social computing in decision-making should be free from the factor of human incompetence and subjectivity.

The primacy of the relationship, rather than the components, provides a new technology for recognition based on the use of distances between the parameters of processes or phenomena. It does not matter that the sides of the rectangle have absolute values of 5 and 10 centimeters. The main thing is that their relationship with each other is 0.5.

Corruption is a system of immoral relations between citizens within the state. Citizens are only carriers of corrupt relations, which are legalized by the constitution, history and traditions. The fight against carriers of corruption is folly, as a demonstration of the ignorance of the political elite. To defeat corruption means to destroy the laws, which encourage and allow corruption, create a constitution of moral relations and introduce it into the minds of citizens.

Harmonious or moral relations based on the combination of languages, histories, religions, cultures, traditions are the primary cause of success in government, science, education, economics and the standard of living of citizens. "Unity in diversity" is US National Doctrine (In God We Trust). The weakness of the state in the unitarity of relations created and maintained by the ignorant political elite, which separates citizens for subsequent opportunity and ease of their subsequent destruction.

Creative and tolerant relations between people are achieved primarily by the harmony of the diversity of linguistic cultures, which is the main argument in creating a prosperous state. The monopoly of unitary linguistic and historical culture forms the relations of Nazism and racism, leading to the self-destruction of the state.

Gathering talented people and making them work effectively in a team is the main quality of a professional manager. To expel anyone who is smarter is the slogan of incompetent leaders of all levels who do not have development goals.

V. METRICS, AXIOMS AND THE EVOLUTION OF SOCIAL RELATIONS

Definitions of computing, useful for understanding and practical use: 1) Computing is a process of purposeful

development of the components involved in it. Computing is an interactive relationship between control and execution mechanisms. 2) Everything in the world is computing and nothing else. 3) The simplest types of computing available for understanding and implementation are the following: reading-writing, speaking-listening, monitoring-control. 4) All processes in nature are determined and focused. Chaos and probability, as a phenomenon, is a product of incompetent computing, or a fig leaf on the naked body of our ignorance, according to Einstein. 5) The relationship that generates the elements involved in it is the primary one. There are no elements without relationship. None of the derivative components of the process can exist independently. 6) The process that gives rise to phenomena or components interacting into it is the primary one. The process or computing of the interaction of chicken and egg is the primary. 7) Chicken and egg phenomena are derived from computing or the evolutionary process. 8) Evolution according to Darwin is computing of natural phenomena in time and space. 9) Social computing is the process of developing social relations between the political elite and citizens in time and space to achieve their goals.

VI. METRICS OF RELATIONSHIPS

The elementary basis of the Universe is the relationship between two components: processes or phenomena. As a rule, for process it is the ratio of the inequality of a pair of components (control-execution), which is measured by the equality relation (xor, not-xor), which is the essence of the metric.

There is no single component without relationship, since one component cannot be measured, and since this procedure is a relationship between two components. This is also true in the case when the component itself is in relation of reflexivity with itself. Therefore, the element is defined and considered as part of the relationship.

The metric of the primacy of the relationship of unequal components extends globally to all phenomena and processes, explains them and gives rise to them: 1) Unit - Zero. 2) Black - White. 3) Rich - Poor. 4) Good and Evil. 5) Alphabet - Signature. 6) Student - Teacher. 7) Leader - Executive. 8) Element - Set. 9) Man - Woman. 10) Monitoring - Management. 11) Chicken - Egg. 12) Space - Time. 13) Matter - Energy. 14) Reading - Writing. 15) Listening - Speaking. 16) Process - Phenomenon. 17) Chaos and Order. 18) Elite and the People. 19) Living - Inanimate. 20) Cyber- and Physical-Space.

The interaction of opposite asymmetric phenomena in time creates a stable structure and process of evolution. The interaction of synonymous unitary phenomena in time creates an unstable structure and process of system degradation.

Naturally, in society, the main and primary is the relationship, the secondary is the political elite and who exist only through their relations in the process of evolution.

VII. ASYMMETRY AND EVOLUTION OF RELATIONSHIPS

Axiom 1. Derivative or symmetric difference by opposite phenomena or processes is equal to their union. The derivative with respect to all 20 mentioned pairs of relations is equal to

their union. The derivative with respect to synonymous phenomena or processes is zero.

Axiom 2. Evolutionary relations always create asymmetry or inequality of interacting components. Matter and antimatter. Relations between them at the level of mesons, which were more, and antimesons created the Universe as a result of annihilation.

Axiom 3. Equivalence of unitary components of the relation is not capable to evolving. Therefore, the ratio of equality of components in the system means the end of development. The criterion for this fact is the zero value of the derivative between the interacting components of the system.

Axiom 4. The components of the relationship in the process of system evolution turn into each other.

Axiom 5. Relation generates elements, but not vice versa.

Axiom 6. The goal of the process is an evolutionary transition from one phenomenon to another. Matter is transformed into energy, time into space, lies into truth, subordinate to leader, egg to chicken, ignorance to knowledge, man to a monster, and vice versa.

The cooperation of equally intelligent students, citizens, workers, scientists, companies, countries, even theoretically does not make sense. Equally minded citizens with equivalent knowledge form the conditions for the degradation of a social group, company, university, or state. Ukraine and Russia will rise together to unprecedented economic heights. Is it profitable for Ukraine to go to Europe? Not today. The derivative between them is such that the one-way flow of qualified and best personnel from Ukraine to Europe is increased in exchange for goods and services from the West. Ukraine and Russia: the derivative between them was in favor of Ukraine in the early 90s. The political elite of Ukraine did not take advantage of this, always ignorant in the management, today and 30 years ago.

The outflow of personnel goes towards Russia. Goods, raw materials and services come from Russia. There is only one way out: to create friendly relations of constructive cooperation, change the constitution and legislation in order to attract specialists, goods and services from Russia and to Russia in equal proportions. Obviously, the derivative with respect to the intelligence of the same people is zero. The gender politics of equality a priori of unequal sexes are not stupid, but the deliberate destruction of humanity through artificial gender equality. Marital relations of relatives with the same genes, the derivative between them is zero, are also doomed to degradation.

Development or evolution is always a consequence of the asymmetry (inequality) of relations, where the derivative between the components has a maximum value equal to their union. Here are some illustrative historical examples of equality of relationships. Communism is the practice of the theoretically false doctrine of equality of social relations, which led to the stagnation and disintegration of statehood. Equality of scientist salary in universities leads to the systematic destruction of science and education in the country. Symmetry or equality of relationships means the lack of development of a scalable social group. Equality of relations and the degradation of society are strictly correlated concepts. Functions and, or, not are asymmetric and create evolution. What are the functions xor, not-xor, which are strictly symmetric equality relations?

Answer: only to measure asymmetries or asymmetrical relations of matter and energy in space and time!

Conclusion: a self-developing or evolving system should have asymmetrical unequal components and provide at each moment of time a derivative between them that does not equal to zero (it is maximum, equal to the union of components), which is measured by symmetric (xor, not-xor) functions. The stability (power) of the system depends on the presence of opposite parts, which have the maximum distance between them (according to Hamming); this makes it possible to obtain the maximum value of the xor operation or a symmetric difference equal to the combination of non-intersecting components.

An example of this is the university's social system, which has no intersection in employees between management mechanisms (academic council) and execution (rector's office), which makes it possible to obtain the maximum stability of the system (power of interacting components) equal to their combination. The farther apart the components, the more powerful or stable the relationship, system, phenomenon or process. A system with the same components, which create a relationship is not able to function. The stability or power of the system is determined by the symmetric difference or Hamming code distance between the interacting components. The given estimate of power or stability is determined by the distance function or the ratio of the number of unit values in the resulting vector of xor-interaction of a pair of components related to the total number of coordinates.

The symmetric difference of the procedural components in each phase of the system development is equal to the universal unit. The symmetric difference of components with respect to opposites is equal to the universal unit. This means that the xor-interaction of two components of a stable system at each time instant is equal to the unit vector in all its coordinates. An example is the system of sustainable development of the Universe, if we consider the pair: matter – energy, space – time.

The xor-interaction (comparison) of a pair of asymmetric functions and, or (and-not, or-not) gives a symmetric function xor: (0001) xor (0111) = (0110), (1110) xor (1000) = (0110).

The violation of symmetry through the occurring defects leads to the detection of single faults.

The efficiency metric from the Creator operates with four components, which are mutually transforming into each other: matter-energy, space-time. The effectiveness metric developed by human operates with three components: time, money, quality (in the metric: matter-energy, space-time). Here, the last two elements are the product of social relations that have arisen as a result of mistrust of human activities violating physical and social laws and norms.

The creator is free from such violations, since everything he does is related to the determinism of natural laws (gravity, reaction, speed of light), which cannot be broken. To go with the analogy of the development of nature means to create and adopt such laws, which cannot be broken. All other is human ignorance, which must be destructed. There are non-enforceable laws in society, which are passed by ignorant people.

A law that is not enforced is not a law. Deputies must be held financially responsible for such lawmaking. The adoption of the

law is preceded by in-depth studies on the conditions and infrastructure of its implementation, provision, execution with mandatory computer simulation of the consequences of its implementation in society. If the law can even be theoretically broken, it should not be passed. At least 90 percent of the laws can be destructed, which will lead to an increase in the activity of citizens and the quality of their life. If the number of prohibitions and restrictions is significantly more permits, then it is simpler to indicate to a citizen a minimum set of laws defining what is not prohibited. This is the mathematical essence of defining the functions of legislation of direct action for each citizen.

VIII. CONCLUSION

1) Social computing is the moral digital and direct human-free management of every citizen based on accurate monitoring of his/her preferences, which makes it possible to destroy corruption, improve the quality of life of citizens and preserve the ecology of the planet.

2) A person, as incapable of effectively managing himself without mistakes, must trust in cyber social computing, which will save humanity from its vices, direct its efforts and mind to the moral solution of the existing problems of energy, materials, ecology, quality of life.

3) Solving the problems of social group management should be viewed from the point of deterministic metric computing, which will gradually replace management based on statistical analysis that does not take into account the interests of each individual citizen.

4) Democracy, as an unscientific method of probabilistic management of social groups and making incompetent decisions, should be replaced by the moral cyber-computing of metric management of society based on exhaustive testing of the interests of every citizen.

5) To unite social groups to effectively address economic and industrial problems is possible only through the adoption of a constitution and laws, which integrate citizens' efforts based on the doctrine of the unity of diversity of languages, religions, histories, traditions and cultures.

REFERENCES

- [1] D.C. Tarraf, "Control of Cyber-Physical Systems," Workshop held at Johns Hopkins University, March 2013, Springer, 2013.
- [2] M.A. Khan, H. Debnath, C. Borcea, "Balanced Content Replication in Peer-to-Peer Online Social Networks," 2016 IEEE International Conferences on Big Data and Cloud Computing, Social Computing and Networking, 2016, pp. 274-283.
- [3] M. R. Lee, T. T. Chen, "Understanding Social Computing Research," IT Professional, 2013, vol. 15, iss. 6, pp. 56-62.
- [4] J. Higg, V. Gurupur, M. Tanik, "A Transformative Software Development Framework: Reflecting the paradigm shift in social computing," 2011 Proceedings of IEEE Southeastcon, 2011, pp. 339-344.
- [5] V. Hahanov, S. Chumachenko, E. Litvinova, A. Hahanova, "Cyber-physical social monitoring and governance for the state structures," 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), 2018, pp. 123-129.
- [6] V. Hahanov, "Cyber Physical Computing for IoT-driven Service," New York, Springer, 2018.

Forest Areas Segmentation on Aerial Images by Deep Learning

Vladimir Khryashchev
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
v.khryashchev@uniyar.ac.ru

Anna Ostrovskaya
People's Friendship University of Russia
Moscow, Russia
ostrovskaya_aa@rudn.university

Vladimir Pavlov
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
i@yajon.ru

Roman Larionov
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
r.larionov@uniyar.ac.ru

Abstract — The aim of this research is to create a deep learning algorithm for automated forest areas segmentation on high-resolution aerial images. Loss values and Dice coefficient, which compares results of algorithms with real masks, was used to measure the quality of developed model. On the whole this paper demonstrates how convolutional neural network implemented on modern GPUs can be applied for the detection of forests on satellite images.

Keywords — *computer vision, image segmentation, aerial images, forest area detection.*

I. INTRODUCTION

Remote sensing cannot completely replace ground-based data collection, but can help in monitoring large areas and hard-to-reach regions. Although computer vision algorithms for detecting objects in image develop quickly methods of automatic segmentation of satellite images is inferior to the man by the quality of work. Despite the rapid development of computer vision algorithms for detecting objects in an image, the task of segmentation of images of remote sensing of the Earth's surface has not been brought to automatism with similar accuracy as with manual marking [1]. Although a human is able to solve a segmentation problem better than a computer, it takes too much time. In addition, in this case it is impossible to obtain results in real time, so the task of satellite image segmentation using computer vision algorithms is particularly relevant.

Today, large number of algorithms for detecting objects in an image exist. This problem is solved in the biometrics, medicine and robotics [2]. Most of these algorithms can be applied to remote sensing tasks. Segmentation of satellite images is a difficult task. The main approach of its solution based on machine learning methods is marking of the image pixels to corresponding classes of objects. Nowadays, the greatest effectiveness of solving this problem is achieved by using convolutional neural networks. The uniqueness of this method is based on the automatic determination of descriptors in the training process. So improvement of the segmentation accuracy and unique features that distinguish one class of objects from others is achieved [3].

Moreover, there are some reasons that CNN isn't trivial solution of image segmentation. A unique approach is

needed to solve the problem of the spatial extent of detected objects, taking into account the invariance to rotation of image or rescaling [4]. Such algorithms should [5, 6]:

- Have a sufficient number of sample images of each class in the training set. As a rule, open satellite image datasets don't contain enough images. It's necessary to expand the training sets of images by self-marking and mixing images from several datasets.
- Be invariant to rotation. Objects in the image can be rotated absolutely at any angle. The segmentation algorithm should be able to select the borders, regardless of the positioning of object in the image.
- Capture small spatial extent of objects. Most neural network algorithms solve the problem of selecting a large object in the image. These objects can be found in the ImageNet database [7]. Satellite images have a high resolution and cover large area where you need to find small, compared to the whole scene, objects. At the same time, if the work is not done with images of centimeters/pixel, most classes are deprived of unique small details that could become good distinguishing features of the class.

This paper discusses satellite imagery of the forest surface, which has the following features:

- high repeatability of shooting, thanks to which it is possible to repeatedly obtain data on the territory of interest, which increases the probability of obtaining cloudless or low-cloud images;
- large surface area with high spatial resolution;
- multichannel multispectral photos in the visible and invisible ranges, including ultraviolet and infrared channels

Information obtained in the process of shooting with RGB-channels, as well as with the near infrared channel (NIR), has a number of features in terms of their use in the analysis of forest areas.

The blue zone of spectral radiation is actively absorbed by chlorophyll (mainly chlorophyll B). This area is very sensitive to atmospheric conditions such as fog or haze. Compared with red or NIR channels, blue is less sensitive to changes in chlorophyll content. As a result, it is used only for special purposes, for example, water monitoring. To solve the problems of forestry, it's best to use green and red channel composites to obtain high-quality color images that serve as the basis in geographic information systems. Blue channel facilitates recognition of forest fires in cloudless images [8].

Healthy vegetation mostly absorbs in red and blue spectrum, reflecting much of the green. The green channel serves not only to form a composite RGB image, but also allows humans to classify vegetation when it's used in combination with other spectral channels. It's also indispensable in assessing the overall condition of the forest

The red channel is very important for the analysis of vegetation (mainly forests) and is actively used. The wavelength of the red channel is greater than the blue one. For this reason, the state of the atmosphere affects it much less. The red channel plays a crucial role in the analysis of changes in forest cover, for example, in the mapping of damage from natural disasters, classification of vegetation species, monitoring of forest cover, etc. [9].

The reflectivity of tree foliage varies greatly in different species. The reflecting capacity of coniferous leaves is much lower than that of deciduous ones. Values (NDVIRE) NIR of young coniferous forest is higher than old one. Therefore, the NIR channel is very important for forest classification, determination of species composition, as well as for monitoring forest infestation by pests. The NIR channel also plays a key role in mapping the effects of hurricane winds, and is now becoming an important component in the calculation of some indicators that determine the biophysical parameters of vegetation.

This paper presents convolutional neural networks that can be used for forest segmentation. The training process,

testing and special metrics for assessing the quality of neural network work are described.

II. NEURAL NETWORK ARCHITECTURE

In this section we describe architecture of the neural network which was used for segmentation of forests in images taken by two different satellite sources.

The network is based on very popular and widespread architecture called U-Net, which is a convolutional neural network used for the task of semantic segmentation. Originally it was developed for segmentation of medical images of neuronal and other structures and outperformed many other competitors on the ISBI challenge [10].

U-Net is a u-shaped convolutional network, that is it consists of two parts: encoder and decoder. Both encoder and decoder are CNNs consisting of six blocks. Encoder's each block includes two convolutions followed by rectified linear unit activation (ReLU) and max pooling operation. Encoder represents downsampling path. Decoder also consists of six blocks where each block includes up-convolution to upsample spatial size of each map, concatenation with corresponding feature map from downsampling path and two convolutions followed by ReLU. Decoder represents upsampling path which is used for restoration of segmentation mask. The last layer of the network is a convolutional K-channel layer, where K is the number of classes and its output is computed by applying pixel-wise softmax function. In our task, K is equal 2.

Since our images contain four channels: regular RGB and near-IR (NIR) channel, we modified U-Net by adding another encoder which separately accepts processes NIR channel. Results of both encoders are concatenated in the center of the network and also in each block of upsampling path like in original U-Net (see Fig. 1).

Modified version of the network was implemented in Python language using Tensorflow library. Tensorflow is a high performance graph-computation library leveraging GPUs for fast numerical computation and used for machine learning and deep learning tasks.

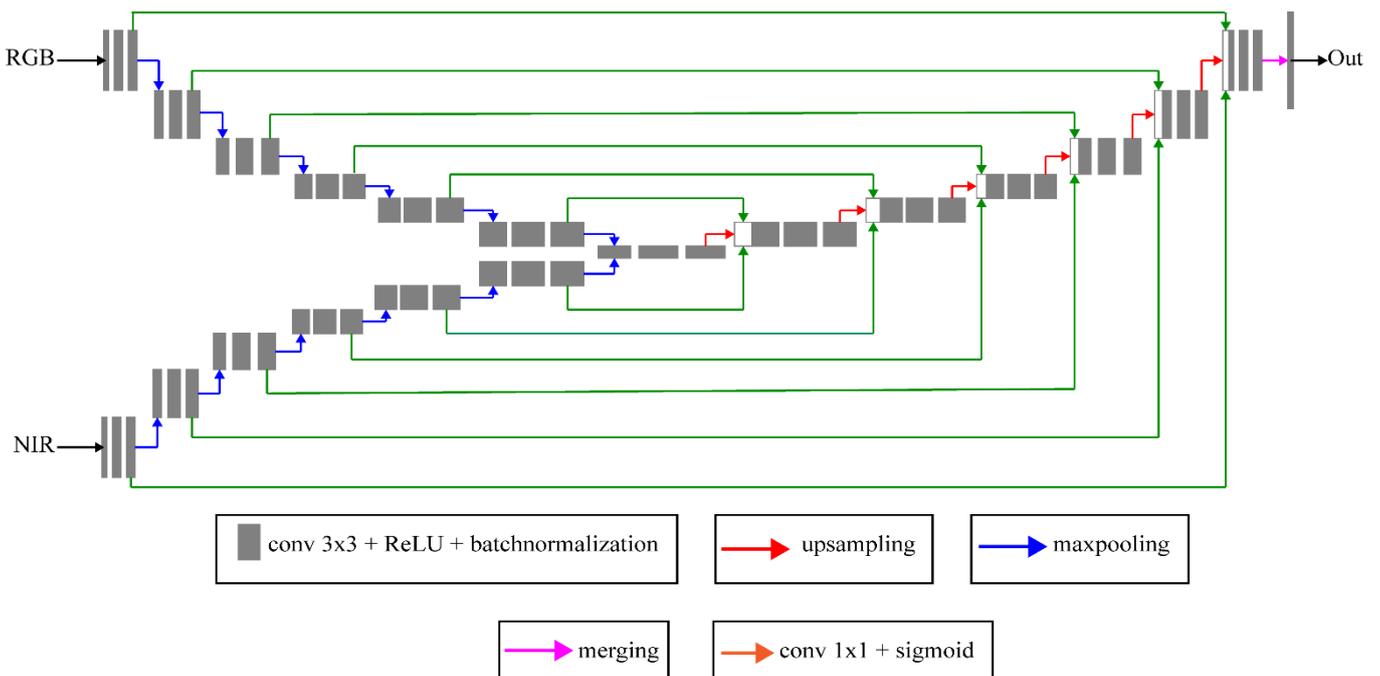


Fig. 1. Architecture of U-Net network with 2 encoders

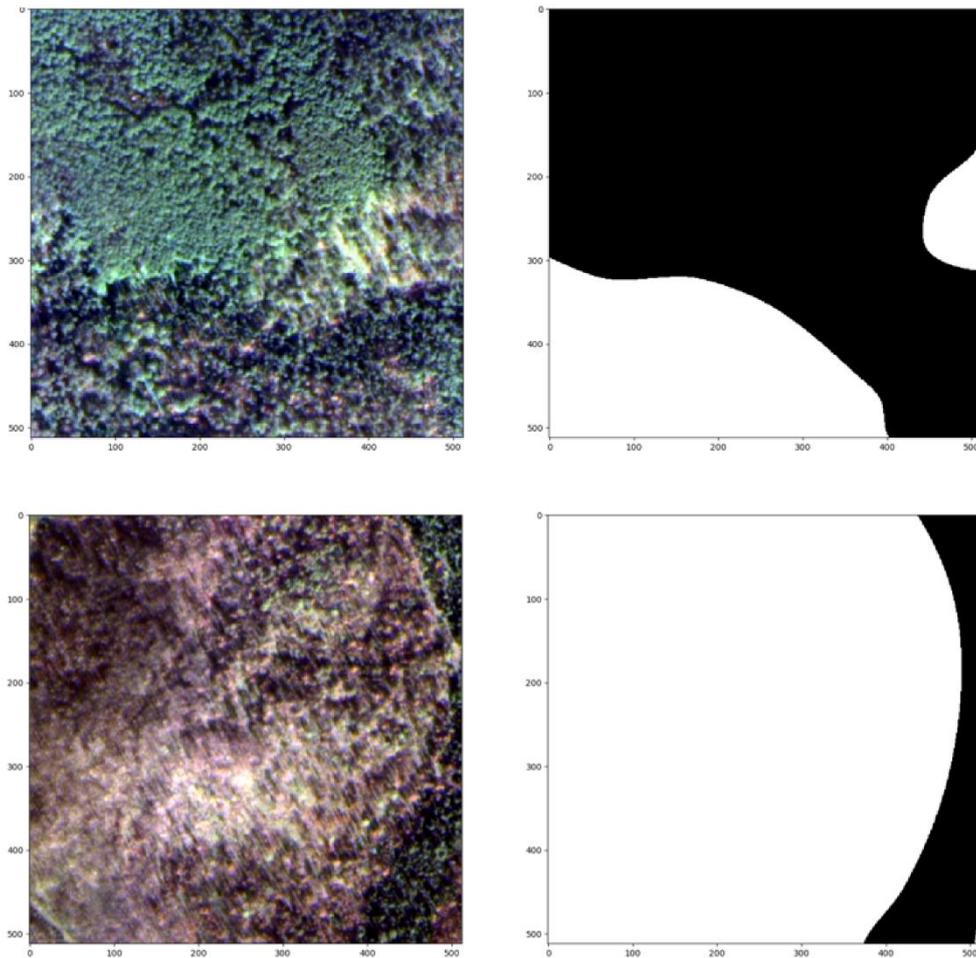


Fig. 2. Examples of extracted RGB-patches and its corresponding masks.

III. NUMERICAL RESULTS

It was necessary to perform some preprocessing steps before directly training the network. The whole dataset consists of 17 images in total from and each image channel contains 16-bit values unlike 8-bit RGB and also prone to the problem of outliers (or impulse noise). Presence of outliers negatively affects results and overall training time and convergence. So first, to tackle this problem we performed per-channel histogram equalization with min-max values chosen by thresholding cumulative distribution function of channel intensities (also known as so-called cumulative count cut). After this preprocessing step each image channel contains values in $[0, 1]$ range with equalized histogram.

Since the dataset is very small, the second step was to split images into training and testing sets, which was performed by cutting each original image horizontally and taking part containing approximately 15% of objects' pixels into test set, while taking the rest into training set.

The third step consisted in extracting patches from training and testing set as images sizes vary from 700 pixels to 12000 pixels per side and it would be impossible to train network using such data due to limited amount of GPU memory. Each patch represents image of size 512 by 512 pixels extracted from original image using sliding window with step 256 pixel per each axis. Examples of extracted RGB-patches and corresponding masks are shown at Fig. 2. Training set was also split into smaller training set and validation set to compute metrics to control

training process per each epoch. Sizes of training and testing sets can be seen in Table I.

TABLE I. SIZES OF TRAINING AND TESTING SETS

Training set	
Total number of patches	14139
Number of patches with objects	8655
Number of patches without objects	5484
Testing set	
Total number of patches	3282
Number of patches with objects	1785
Number of patches without objects	1497

It can also be seen that training dataset is imbalanced what negatively affects network's learning process. To reduce this imbalance each batch is constructed by randomly selecting $N/2$ patches without objects and the same number of patches with objects. This way of constructing batch showed better segmentation and detection results in comparison to standard sequential feeding of images to network.

Moreover, to further increase size of training set we add different image augmentations: random flips (RF); rotations, spatial shifts, shifts in scale (SSR) and random noise in HSV color scheme (RN) which significantly increases quality of final segmentation.

Learning process is monitored after each epoch by evaluating loss and Dice coefficient computed on validation set. Dice coefficient:

$$D(X, Y) = \frac{2|X \cap Y|}{|X \cup Y|}$$

where X and Y are grayscale or binary masks. We used two losses to train models:

1. Binary cross entropy loss + dice loss (BCE + DL):

$$L = \sum_{x,y} BCE(p_x, y) + 1 - D(x, y),$$

$$BCE(p_x, y) = -\log p_x$$

where

$$p_x = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{otherwise} \end{cases}$$

y is the label of sample x , p is the predicted probability of sample x and D is the Dice coefficient.

2. Focal loss + dice loss (FL + DL):

$$L = FL + 1 - D$$

where

$$FL(p_x) = -(1 - p_x)^c \log p_x$$

is the focal loss and $c \geq 0$.

Focal loss was specifically developed for tasks where the data is highly imbalanced and outperformed other competitors on COCO dataset.

We trained the following four models (BB is for balanced batch) during 100 epochs with batch size of 16 images and Adam optimizer:

1. BCE + DL / BB / RF + SSR (BCE #1)
2. BCE + DL / BB / RF + SSR + RN (BCE #2)
3. FL + DL / RF + SSR (FL #1)
4. FL + DL / RF + SSR + HSV (FL #2)

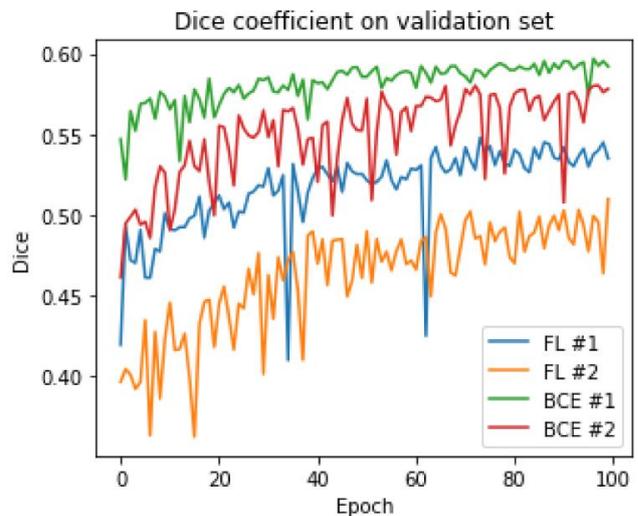
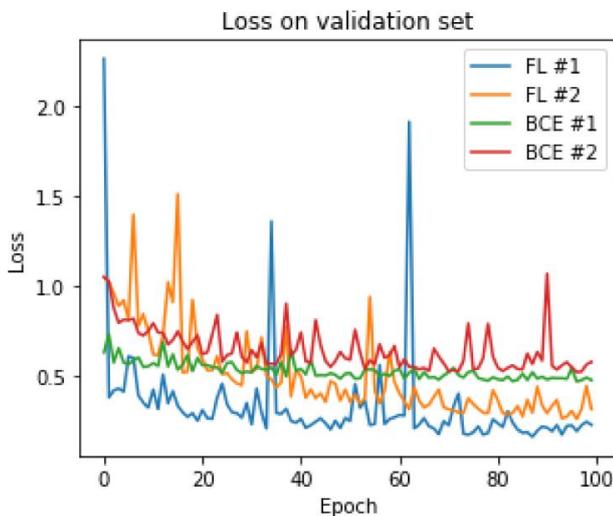


Fig. 3. Loss values and Dice coefficient on computed validation set

Models were trained on supercomputer NVIDIA DGX-1 in AI-center of P.G. Demidov Yaroslavl State University. Loss values and Dice coefficient values changes during validation process are shown in Fig. 3.

The final network output is the segmentation map with values ranging from 0 to 1 and it is necessary to select appropriate value for threshold T to obtain binary mask. For this task we used simple grid search to maximize F1 on validation set with step of 0.01. Selected values of threshold for different models are shown in Table II.

TABLE II. SELECTED VALUES FOR THRESHOLD ON VALIDATION SET

Model	T	F1
BCE #1	0.39	0.3089
BCE #2	0.62	0.4504
FL #1	0.52	0.2175
FL #2	0.56	0.361

Results obtained on test set are shown in Table III. F1 is the standard F-measure computed using precision P and recall R , both of which are calculated according to correct detection. Object is considered correctly detected if it has $IoU \geq 0.5$ with ground truth object. It can be seen that best results showed model BCE #2 for $T = 0.62$. FL #2 with RN did not show any significant increase in quality of segmentation. Moreover, it can be supposed that it produces too many false positives as P value is quite low. The same can be said about BCE #1.

TABLE III. SEGMENTATION RESULTS ON TEST SET FOR DIFFERENT VALUES OF THRESHOLD

Model	T	D	F1	P	R
BCE #1	0.39	0.8439	0.1641	0.09475	0.6136
BCE #1	0.5	0.8446	0.17	0.09873	0.6105
BCE #2	0.62	0.7652	0.3488	0.248	0.5876
BCE #2	0.5	0.7554	0.3319	0.2309	0.59
FL #1	0.52	0.8222	0.1967	0.1168	0.6211
FL #1	0.5	0.8181	0.1775	0.1038	0.6144
FL #2	0.56	0.7874	0.1549	0.09027	0.5443
FL #2	0.5	0.7632	0.1288	0.07269	0.5664

Examples of final segmentation are shown in Fig. 4. The first image is source patch, the second image is ground truth mask and the third one is algorithm prediction.

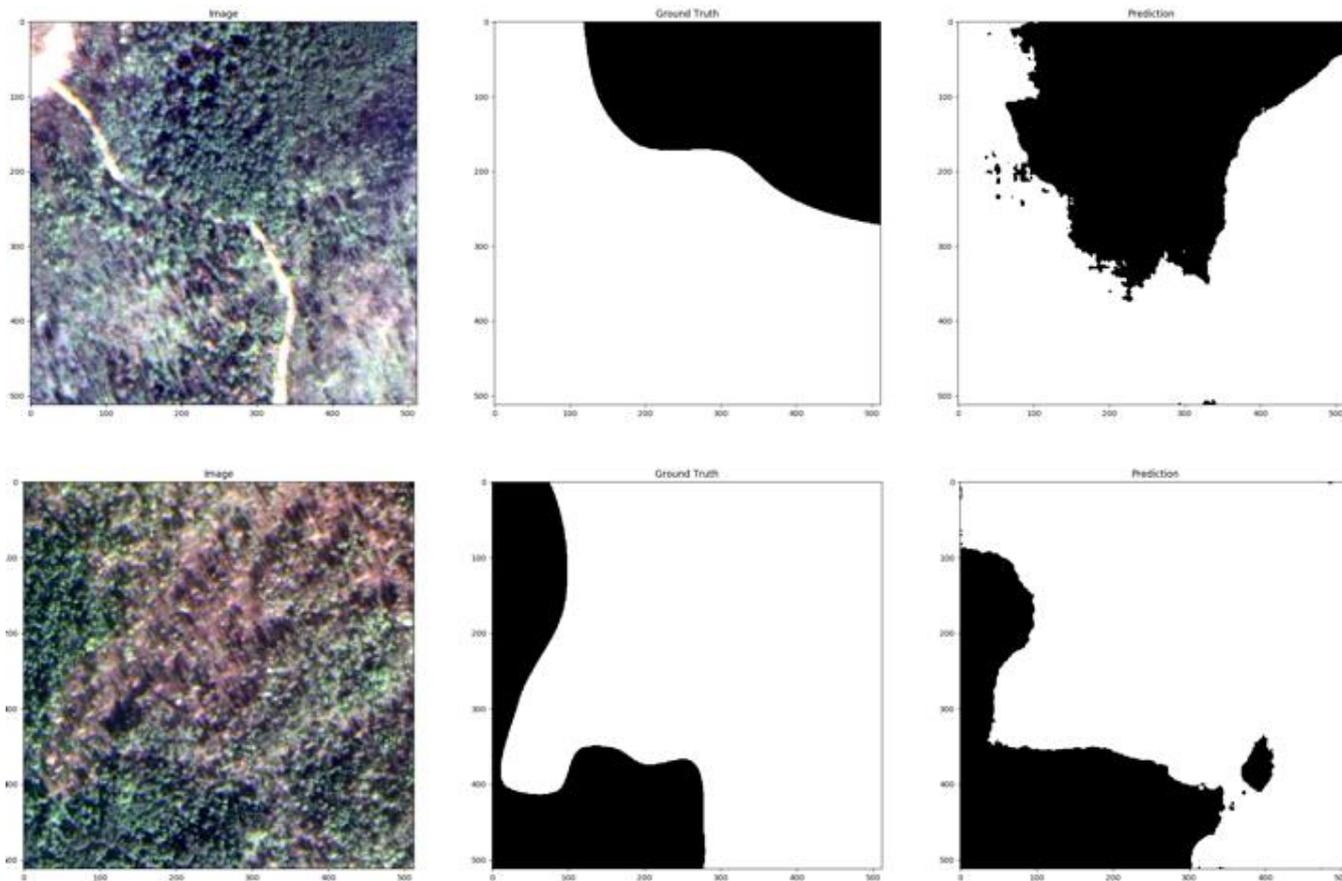


Fig. 4. Examples of aerial forest areas images segmentation

IV. CONCLUSION

In this article we showed that our proposed modification of U-Net with balanced batch is capable of segmenting forests in satellite images and generalization to test set. Although it does not show high results in detection (max value of F1 is 0.34), it can be connected with incorrect or incomplete segmentation of ground truth images by experts, since visually prediction on many test examples look correct. Further research is required to increase quality of segmentation and detection what can be done by increasing number of source images in dataset, adding more aggressive augmentations and usage of modified versions of U-Net with new losses that can correctly tackle problem of imbalanced data.

ACKNOWLEDGMENTS

This article was prepared with the financial support of the Ministry of science and education of the Russian Federation under the agreement No. 075-15-2019-249 from 04.06.2019 (identifier works RFMEFI57517X0167).

REFERENCES

- [1]. Sayfeddin D., Bulgkov A., Kruglova T. Neurosetevaya sistema otslejvaniya mestopolozheniya dinamicheskogo agenta na baze kvadrokoptera // Inzhenernyi vestnik Dona, 2014, No. 1, URL: <http://ivdon.ru/magazine/archive/n1y2014/2293/>
- [2]. Plutogarenko N., Varnavskiy A. Primenenie neironnykh setei dlya postroeniya modeli prognozirovaniya sostoyaniya gorodskoi vozduшной sredy // Inzhenernyi vestnik Dona,

2012, No. 4-2, URL: <http://ivdon.ru/magazine/archive/n4p2y2012/1351/>

- [3]. L. Deng, D. Yu Deep Learning : Methods and Application Foundations and Trends in Signal Processing. 2014, vol. 7, no. 3-4, pp. 197-387.
- [4]. Solov'ev R. A., Telpukhov D. V., Kustov A. G., Avtomaticheskaya segmentatsiya sputnikovykh snimkov na baze modifitsirovannoi svertochnoi neironnoi seti UNET (Automatic segmentation of satellite images based on the modified UNET convolutional neural network), Inzhenernyi vestnik Dona, 2017, Vol. 47, No. 4(47), URL: http://ivdon.ru/uploads/article/pdf/IVD_56_soloviev_N.pdf_116222c2f5.pdf.
- [5]. V.Khryashchev, L.Ivanovsky, V.Pavlov, A.Ostrovskaya, A.Rubtsov Comparison of Different Convolutional Neural Network Architectures for Satellite Image Segmentation // Proceedings of the FRUCT'23, Bologna, Italy, 13-16 November 2018. pp.172-179.
- [6]. Khryashchev V., Pavlov V., Priorov A., Ostrovskaya A. Deep learning for region detection in high-resolution aerial images // IEEE East-West Design & Test Symposium (EWDTS). 2018. pp. 1-5.
- [7]. ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), Web: <http://image-net.org/challenges/LSVRC/2012/>
- [8]. W G. Rees Physical principles of Remote Sensing. Cambridge University Press, 2006, 336 p.
- [9]. Campell J.B. Introduction to remote sensing / J.B. Campell.- N.Y.-London: The Guilford press, 1996-P. 120-549.
- [10]. O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, vol. 9351, 2015, pp. 234-241.

An Analysis of LockerGoga Ransomware

Alexander Adamov
NioGuard Security Lab /
Design Automation Dep.
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
ada@nioguard.com /
oleksandr.adamov@nure.ua

Anders Carlsson
Dep. of Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
anders.carlsson@bth.se

Tomasz Surmacz
Institute of Computer Engineering,
Control, and Robotics
Wrocław University of Science and
Technology
Wrocław, Poland
tsurmacz@ict.pwr.wroc.pl

Abstract—This paper contains an analysis of the LockerGoga ransomware that was used in the range of targeted cyberattacks in the first half of 2019 against Norsk Hydra - a world top 5 aluminum manufacturer, as well as the US chemical enterprises Hexion, and Momentive - those companies are only the tip of the iceberg that reported the attack to the public.

The ransomware was executed by attackers from inside a corporate network to encrypt the data on enterprise servers and, thus, taking down the information control systems. The intruders asked for a ransom to release a master key and decryption tool that can be used to decrypt the affected files.

The purpose of the analysis is to find out tactics and techniques used by the LockerGoga ransomware during the cryptolocker attack as well as an encryption model to answer the question if the encrypted files can be decrypted with or without paying a ransom.

The scientific novelty of the paper lies in an analysis methodology that is based on various reverse engineering techniques such as multi-process debugging and using open source code of a cryptographic library to find out a ransomware encryption model.

Keywords—Ransomware, LockerGoga, Malware, Reverse Engineering, Malware Analysis, cryptolocker, encryption, cryptography, targeted attack.

I. INTRODUCTION

In March 2019, BleepingComputer reported that LockerGoga was allegedly responsible for disrupting the Norsk Hydra IT control system and forced the Norwegian industrial giant to switch to the manual operation mode [1]. Later, according to Motherboard, this ransomware disrupted IT services of the US chemical companies Hexion and Momentive [2]. Thus, it seems that the attackers behind LockerGoga target critical infrastructure and those mentioned above are not the only victims of the ransomware up to the moment. Further we provide the detailed analysis of the ransomware encryption process.

II. BYPASSING ANTIVIRUS

Antiviruses missed the LockerGoga sample supposedly because the ransomware had the valid digital signature.

When the sample (SHA256: eda26a1cd 80aac1c4 2cddbba9a f813d9c4 bc81f605 2080bc33 435d1e07 6e75aa0) was firstly uploaded on March 8, 2019 to VirusTotal [3], it had 0 detections out of 67 security products.



Fig. 1. Virustotal verdicts for the LockerGoga ransomware sample

The ransomware version 1320, which is under analysis, has the following digital certificate issued to the fake 'ALISA LTD' entity by Sectigo RSA Code Signing CA, well known for signing abuse according to the Chronicle research [4], but was revoked after discovery of the attack.

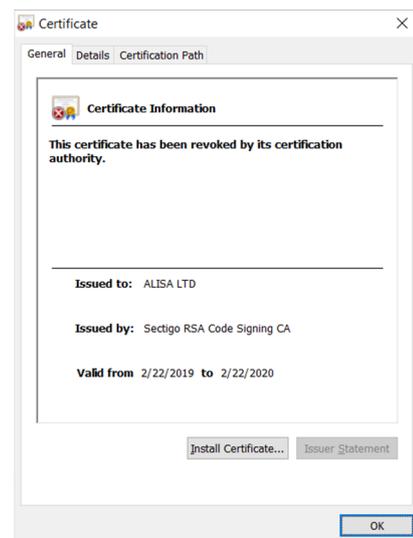


Fig. 2. The LockerGoga's certificate

III. STATIC AND DYNAMIC ANALYSIS

The binary contains statically linked *boost* and *Crypto++* library that complicates the analysis of the ransomware, even though the techniques such as obfuscation, packing, or code encryption have not been used.

Once started, the cryptolocker copies itself to the %Temp% folder under the hardcoded name.

```
C:\Windows\system32\cmd.exe /c move /y  
"C:\LockerGoga ransomware"  
C:\Users\<USER>\AppData\Local\Temp\yxugwjud<ID>  
.exe
```

After that, it executes the master process with the '-m' key.

The master process creates the list of files to be encrypted. The version 1320 of the cryptolocker does not perform filtering of the files based on extensions and encrypts all accessible files on disks.

A. Inter-process communication (IPC)

The master process sends a task to a worker through the named shared memory created with `CreateFileMapping` providing a path to file for encryption. The worker gets access to the master's named shared memory by calling the function `OpenFileMapping` using the identifier 'Global\SM-yxugwjud'.

Then the master process starts workers with the parameter '-s' also providing the identifier of the created named shared memory '-i Global\SM-yxugwjud'.

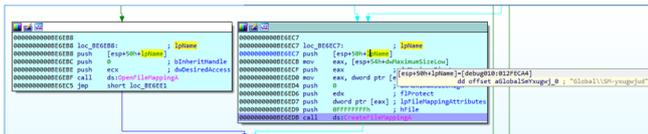


Fig. 3. Using Named Shared Memory for IPC

LockerGoga starts a slave process sending the named shared memory as a command line argument:

```
C:\Users\IEUser\AppData\Local\Temp\yxugwjud1342.exe -i Global\SM-yxugwjud -s
```

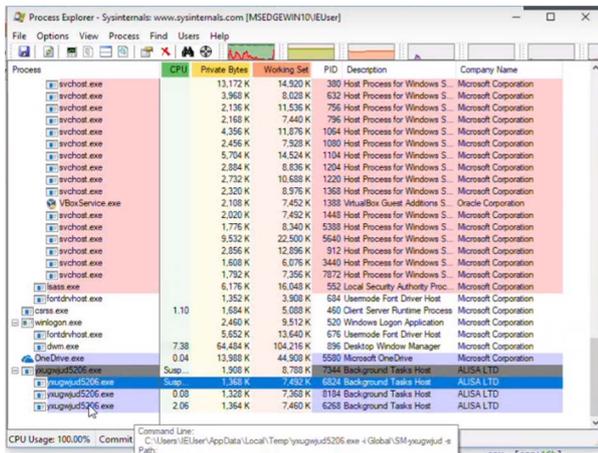


Fig. 4. The LockerGoga processes in ProcessExplorer

B. File operations

The worker tries to open the file by the path given by the master. Then, it requests write permissions for the target file using the `boost` library. It deletes the '.locked' version of the file if it exists before encryption. After that, the worker renames a file to the one with the '.locked' extension and starts encrypting the file content in 65536-byte blocks.

IV. ENCRYPTION MODEL

In this section, we'll find out the encryption algorithms, keys, and modes used to encrypt files and file keys with the help of the debugger and `Crypto++` library open source code [5].

A. File encryption

For file content encryption LockerGoga uses AES in CRT mode and key length of 128 bits.

The file key and Initialization Vector (IV) are generated using the MS Crypto Provider's `CryptGenRandom()` function.

```
00000000C5AF5E push esi
00000000C5AF5F push eax
00000000C5AF60 lea eax, [ebp+var_C]
00000000C5AF63 mov large fs:0, eax
00000000C5AF69 mov esi, [ebp+pbBuffer]
00000000C5AF6C lea ecx, [ebp+var_51]
00000000C5AF6F call sub_C5B0A0
00000000C5AF74 push esi ; pbBuffer
00000000C5AF75 push [ebp+dwLen] ; dwLen
00000000C5AF78 push dword ptr [eax] ; hProv
00000000C5AF7A call ds:CryptGenRandom ; gen AES key and IV
00000000C5AF80 test eax, eax
00000000C5AF82 jz short loc_C5AFA0
```

Fig. 5. CryptGenRandom call to generate AES and IV.

The encryption mode of AES can be found out using the following algorithm:

1. Find the crypto-related string constants in Assembly code of the statically linked crypto library.

Address	Length	Type	String
.rdata:00CC...	000...	C	: this object requires an IV
.rdata:00CC...	000...	C	: this object cannot use a null IV
.rdata:00CC...	000...	C	is less than the minimum of
.rdata:00CC...	000...	C	: IV length
.rdata:00CC...	000...	C	exceeds the maximum of

2. Locate the corresponding high-level C++ open source code by the discovered string constant.

`cryptlib.cpp`:

```
void SimpleKeyingInterface::
ThrowIfResynchronizable() {
    if (IsResynchronizable())
        throw InvalidArgument(
            GetAlgorithm().AlgorithmName()
            + ": this object requires an
            IV");
}
```

3. Study the class structure to figure out what data an object of this cryptography class may contain. In this case, the C++ method `ThrowIfResynchronizable()` of the `SimpleKeyingInterface` class throws an exception with the algorithm name.
4. Set up a breakpoint in the ransomware code or change the execution order by modifying the Instruction Pointer so that encryption info (e.g. the algorithm name) can be seen under the debugger.

```
00000000C3CC44 call dword ptr [eax+1Ch]
00000000C3CC47 cmp eax, 4
```

```
ptr [ebx], 0
00000000C3CC98 loc_C3CC98:
00000000C3CC98 mov eax, [esi]
00000000C3CC9A mov ecx, esi
00000000C3CC9C call dword ptr [eax+34h]
00000000C3CC9F mov ecx, [eax]
00000000C3CCA1 mov edx, [ecx+8]
00000000C3CC44 lea ecx, [ebp+var_54]
00000000C3CC47 push ecx
00000000C3CC48 mov ecx, eax
00000000C3CCAA call edx
00000000C3CCAC push offset aThisObjectRequ ; ": this object requires an IV"
00000000C3CCB1 push eax
```

5. The pointer to algorithm name string is stored in EAX register and shows "AES/CTR".

```
push ecx
mov ecx, eax
call edx
push dword ptr [eax+1Ch]
push eax
lea eax, [ebp+var_6C]
mov byte ptr [eax], 41h ; A
push eax
call sub_C5B0A0
add esp, 4
push eax
lea ecx, [ebp+var_54]
mov byte ptr [eax], 2Fh ; /
call sub_C5B0A0
push offs
```

Then, the AES key and IV are encrypted using RSA 1024-bit public key and stored later in the footer of the encrypted file.

B. File keys encryption

Similarly, we can find the algorithm used for file keys encryption.

The `Encrypt()` interface in `Crypto++` library that may throw an exception with an algorithm name can be analysed:

```
void TF_EcryptorBase::Encrypt(
    RandomNumberGenerator &rng, const byte
    *plaintext, size_t plaintextLength, byte
    *ciphertext, const NameValuePairs
    &parameters) const {
    if (plaintextLength >
        FixedMaxPlaintextLength()) {
        if (FixedMaxPlaintextLength() < 1)
            throw InvalidArgument(AlgorithmName()
                + ": this key is too short to encrypt
                any messages");
        else
            throw InvalidArgument(AlgorithmName()
                + ": message length of " +
                IntToString(plaintextLength) +
                " exceeds the maximum of " +
                IntToString(FixedMaxPlaintextLength())
                + " for this public key");
    }
    SecByteBlock addedBlock(
        PaddedBlockByteLength());
    GetMessageEncodingInterface().Pad(rng,
        plaintext, plaintextLength, paddedBlock,
        PaddedBlockBitLength(), parameters);
    GetTrapdoorFunctionInterface().
        ApplyRandomizedFunction(rng,
        Integer(paddedBlock, paddedBlock.size())
        ).Encode(ciphertext,
        FixedCiphertextLength());
}
```

The algorithm name obtained in this way is RSA/OAEP-MGF1(SHA-1).

```
lea ecx, [esi+4]
call ecx, [edi+1Ch]
push offset aMessageLength0 ; ": message length of "
push eax
lea eax, [ebp+var_c0]
mov byte ptr [eax], 0
push eax
call sub_401000 ; dd offset aRsaOaepMgf1Sha ; "RSA/OAEP-MGF1(SHA-1)"
```

Fig. 6. RSA algorithm name

During the initialization phase, the worker instantiates an RSAFunction object and loads the hardcoded public key.

```
00000000C59F85 push 0Bh
00000000C59F87 push offset aThisObject ; "ThisObject:"
00000000C59F8C push [ebp+var_1C]
00000000C59F91 call sub_C59F5B
00000000C59F94 add esp, 0Ch [ebp+var_1C]=debug008:004FF004
00000000C59F97 test eax, eax
00000000C59F99 jnz short loc_C5A03F ; dd offset aThisObjectClass ; "ThisObject: class GijWRdG: RSAFunction"
```

Fig. 7. RSAFunction object instantiation declared in the Crypto++ library to load the hardcoded RSA public key

```
00000000C5A03F loc_C5A03F:
00000000C5A03F sub esp, 0Ch
00000000C5A042 mov dword ptr [ebp+var_C], offset sub_C224B0
00000000C5A049 mov eax, esp
00000000C5A04B mov dword ptr [ebp+var_C+4], 0FFFFFFF8h
00000000C5A052 movq xmm0, [ebp+var_C]
00000000C5A057 lea ecx, [ebp+var_20]
00000000C5A05A push offset aModulus ; "Modulus"
00000000C5A05F movq qword ptr [eax], xmm0
00000000C5A063 mov dword ptr [eax+8], 0
00000000C5A06A call sub_C597E0
00000000C5A06F sub esp, 0Ch
00000000C5A072 mov dword ptr [ebp+var_C], offset sub_C59F30
00000000C5A079 ecx, esp
00000000C5A07B mov dword ptr [ebp+var_C+4], 0FFFFFFF8h
00000000C5A082 movq xmm0, [ebp+var_C]
00000000C5A087 push offset aPublicExponent ; "PublicExponent"
00000000C5A08C movq qword ptr [eax], xmm0
00000000C5A090 mov dword ptr [eax+8], 0
00000000C5A097 mov ecx, eax
00000000C5A099 call sub_C597E0
```

Fig. 8. Loading RSA public key in Crypto++ library

The public key (Modulus and Public exponent) is in the PEM format and hardcoded in the ransomware.

```
00000000C20A88 push 0
00000000C20A8D push offset aHlgdMaGcsqsg1 ; "HlGDhAGCsq551b3DQEBaQUAA4GLDCBhwKgQ04ZArnc1s7Wav4PzdrEVAM2ai"
00000000C20A92 push ecx, [ebp+var_4B]
00000000C20A97 lea ecx, [ebp+var_4B]
00000000C20A9D mov byte ptr [eax], [ebp+var_4B]
00000000C20A9E call sub_401000 ; db "I++yJd: B1102vB5rE03KoaPH1cXV6aGhdJd1PUpR4Udod3W7CagV/89uI72"
00000000C20AA6 push 1
00000000C20AA8 lea ecx, [ebp+var_4B]
00000000C20AAE mov byte ptr [eax], [ebp+var_4B]
00000000C20AB2 push eax
00000000C20AB3 lea ecx, [ebp+var_20] ; db "TOP1ohdvrzzvzj0zhNw1BQC=" ; 8
```

Fig. 9. The public RSA public key stored in the ransomware code

After decoding the RSA public key, we can see the size of the key and public exponent equal to 17.

```
$ openssl rsa -inform PEM -pubin -in pub.key -text -noout
Public-Key: (1024 bit)
Modulus:
00:f8:64:0a:e6:72:2b:3b:bd:66:af:e0:fc:dd:ac:
4b:d6:5b:66:96:23:ef:a3:62:e0:f3:04:b2:35:39:
9b:f4:4a:b1:0e:dc:aa:1a:3c:c8:f5:71:75:7a:6b:
e1:87:76:78:dd:88:f5:29:ad:4d:1d:a1:d2:56:ec:
26:a0:57:ff:3d:58:8e:f6:45:97:55:45:83:d5:5c:
d2:a8:2a:d5:33:14:cd:7a:2a:28:2e:c0:a6:7a:65:
8f:d9:75:00:a0:2e:dc:2b:67:fd:ab:d8:a2:66:6b:
3a:e4:72:d9:50:b3:3e:96:09:c0:84:4c:e3:35:a2:
17:6b:bf:3c:d6:8c:ec:e1:63
Exponent: 17 (0x11)
```

Fig. 10. The decoded public RSA public key

The cryptolocker uses RSA-1024 with the ‘MGF1(SHA-1)’ mask generation function for the OAEP padding scheme to encrypt 40 bytes buffer that contain first 4 zero bytes, 16-byte file IV, 16-byte file key, and the terminating 4-byte string ‘goga’.

```
00000000C3D387 push dword ptr [edi+20h] ; parameters
00000000C3D38A mov ecx, [edi+1Ch]
00000000C3D38D push dword ptr [edi+58h] ; ciphertext
00000000C3D390 push [ebp+var_54] ; plaintextLength = 40 bytes
00000000C3D393 mov eax, [ecx]
00000000C3D395 push ebx ; plaintext
00000000C3D396 push dword ptr [edi+18h] ; RandomNumberGenerator
00000000C3D399 call dword ptr [eax+10] ; RSA_encrypt
00000000C3D39C mov eax, [ebp+var_1C]
00000000C3D39F mov edi, ebx
00000000C3D3A1 mov ecx, [ebp+var_18]
```

Fig. 11. File key data encryption with RSA

Once encrypted, this footer is appended at the end of the encrypted file.

```
01 23 45 67 89 AB CD E F 10 11 12 13 14 15 16 0123456789ABCDEF0123456
003863EA 78 43 83 08 F2 AD 21 BA 28 76 AA 09 55 28 F8 1F 60 78 F4 1D 91 43 37
00386401 80 A7 0C CD 0F C1 04 62 88 80 71 18 77 12 58 F8 F7 A4 80 64 91 88 46
00386418 4B 01 56 8D 3D 08 92 25 AD FE 04 7C 81 84 70 73 05 K2 1A 98 08 5C
0038642F E7 1B FA 61 D7 31 E1 C0 95 0A F4 1D C0 79 03 0C 60 60 60 60
00386446 F3 D2 2A 3F D6 24 8E 8A 11 55 4A 0E 74 10 33 32 30 62 64 38 00 00
0038645D 22 02 0B 0A 31 C1 63 D8 97 87 4F 87 71 51 33 32 30 62 64 38 00 00
00386474 00 00 0A 15 FF 35 01 B3 96 58 66 31 E7 B0 A0 53 60 B8 0D C8 7E 7F 07
0038649B EE 3B 05 48 0B 80 97 62 DB 7D 5F 30 53 B5 0E 1F 0D 1C 0E 4A AB 20 24
003864B2 51 2D B0 59 31 E1 E2 7E 2B 0A 0C 0E 1C 0A 09 12 7E 00 99 99
003864D9 3D FE 04 C3 0A BC 6F 3E 2B 0A 0C 0E 1C 0A 09 12 7E 00 99 99
003864D0 B0 DA 79 A1 01 21 0E 9B E3 14 A0 0B 9F FC D5 53 0F 05 A5 20 7C 28
003864E7 EA CC 5D E3 D8 6B 8C SA 50 07 92 93 8C A5 A3
```

Fig. 12. LockerGoga footer structure

The low public exponent value (e=17) is mitigated by the OAEP randomizing padding scheme that can be identified by a different footer’s ciphertext appearing while a plaintext and public key are the same.

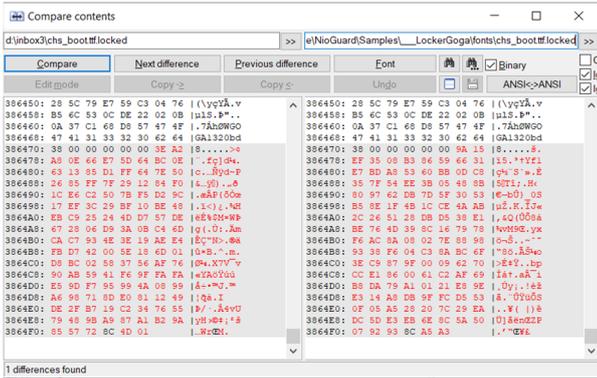


Fig. 13. File key data encrypted with RSA are randomized because of OAEP

To sum up, the discovered LockerGoga encryption model employs symmetric encryption for files (AES-128-CTR) and asymmetric encryption for the file key data (RSA-1024 OAEP/MGF1(SHA-1)) and does not allow a victim to decrypt the locked files without knowing the corresponding RSA private key (the master key).

C. Ransom note

Once files are encrypted, the locker leaves the following ransom note:
 “Greetings!

There was a significant flaw in the security system of your company. You should be thankful that the flaw was exploited by serious people and not some rookies. They would have damaged all of your data by mistake or for fun.

Your files are encrypted with the strongest military algorithms RSA4096 and AES-256. Without our special decoder it is impossible to restore the data. Attempts to restore your data with third party software as Photorec, RannohDecryptor etc. will lead to irreversible destruction of your data.

To confirm our honest intentions. Send us 2-3 different random files and you will get them decrypted. It can be from different computers on your network to be sure that our decoder decrypts everything. Sample files we unlock for free (files should not be related to any kind of backups).

We exclusively have decryption software for your situation

DO NOT RESET OR SHUTDOWN - files may be damaged.
 DO NOT RENAME the encrypted files.
 DO NOT MOVE the encrypted files.

This may lead to the impossibility of recovery of the certain files.

The payment has to be made in Bitcoins.

The final price depends on how fast you contact us.

As soon as we receive the payment you will get the decryption tool and instructions on how to improve your systems security

To get information on the price of the decoder contact us at:
 SuzuMcperson@protonmail.com
 AsuxidOruraep1999@o2.pl”

The ransom note states the different encryption model (RSA-4096 and AES-256) from what has been discovered in this research.

V. DECRYPTION OPPORTUNITY

While the discovered encryption model leaves no chances for decryption without paying a ransom, it is still possible to decrypt the locked files if the cryptolocker is still working. To do that, it is necessary to make dumps of the slave processes responsible for encryption of these files.

Figure 11 shows the file key data located in the memory before being encrypted with RSA and stored at the footer. The buffer contains 4 zero bytes, 16-byte file IV, 16-byte file key, and the terminating 4-byte string “goga”.



Fig. 14. File key data before encryption with RSA can be located in the process memory.

Therefore, it is possible to create a simple Yara rule to scan the process memory or process memory dump for the presence of the key data. An example of such Yara rule [6]:

```
rule LockerGogaInMemKeys :
{
  strings:
    $a = "goga" nocase
  condition:
    $a
}
```

Once the file AES key and IV are found using this method, it is possible to decrypt the locked file (the file path can be also found in the memory dump) with the help of OpenSSL tool [7].

To decrypt an encrypted file for which you have located the key and IV in the memory dump:

1. Make a backup copy of the encrypted file.
2. Delete the 148-byte footer from the encrypted file.
3. Decrypt the file using any cryptographic tool. For example, to decrypt the file encrypted with the key and IV shown on the picture above, run:

```
$ openssl aes-128-ctr -d -in
chs_boot.ttf.locked_nofooter
-K F12D893D2B9E8CC639C2EE3B06617AAC
-iv 44C5A7A5FBF58C0C91D16E075B130070 -out
chs_boot.ttf
```

VI. CONCLUSIONS

The analysis of LockerGoga ransomware presented in this paper revealed the following ransomware tactics and techniques:

- Signing a ransomware with a digital certificate issued to a fake entity.
- Distributing file encryption between processes, so that one worker process encrypts one file only, to bypass an antivirus behavior blocker.

- The discovered encryption model is AES-128-CTR for files and RSA-1024 OAEP/MGF1(SHA-1) for the file key data.
- The locked files cannot be decrypted without knowing the master key (RSA-1024 private key).
- Using the CTR mode of AES for file encryption [8].
- Using low public exponent for RSA that speeds up encryption.
- Using OAEP for RSA [9] that mitigates the weakness of using the low public exponent [10].
- The ransom note states the wrong encryption model.

The encryption model with symmetric and asymmetric encryption algorithms has been first proposed in 1996 by Young and Yung at IEEE Security and Privacy Symposium [11] and implemented in 2005 [12] as a proven extortion model from the cryptographic standpoint that can be potentially used by criminals in future.

REFERENCES

- [1] I. Ilascu, LockerGoga Ransomware Sends Norsk Hydro Into Manual Mode, BleepingComputer, 2019, available at <https://www.bleepingcomputer.com/news/security/lockergoga-ransomware-sends-norsk-hydro-into-manual-mode/>
- [2] L. Franceschi-Bicchierai, Ransomware Forces Two Chemical Companies to Order ‘Hundreds of New Computers’, Motherboard, 2019, available at https://motherboard.vice.com/amp/en_us/article/8xyj7g/ransomware-forces-two-chemical-companies-to-order-hundreds-of-new-computers
- [3] Virustotal service, 2019, available at <https://www.virustotal.com/>
- [4] Abusing Code Signing for Profit. Chronicle, 2019, available at <https://medium.com/@chroniclesec/abusing-code-signing-for-profit-ef80a37b50f4>
- [5] Crypto++ crypto library, 2019, available at <https://www.cryptopp.com/>
- [6] Yara project, 2019, available at <https://github.com/VirusTotal/yara>
- [7] OpenSSL crypto library, 2019, available at <https://www.openssl.org/>
- [8] M. Dworkin, Recommendation for Block Cipher Modes of Operation. Methods and Techniques. NIST Special Publication 800-38A, 2001, available at <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-38a.pdf>
- [9] M. Bellare and P. Rogaway. Optimal asymmetric encryption. In, R. Rueppel editor, Advances in Cryptology - Eurocrypt’94, Lecture Notes in Computer Science, volume 950, pp. 92-111. Springer Verlag, 1994.
- [10] D. Boneh, Twenty Years of Attacks on the RSA Cryptosystem. Stanford University, 1999, available at <https://crypto.stanford.edu/~dabo/pubs/papers/RSA-survey.pdf>
- [11] Young A., Yung M., Cryptovirology: Extortion-based security threats and countermeasures. In Security and Privacy Proceedings, IEEE Symposium, 1996, pp. 129–140.
- [12] Young A. Building a Cryptovirus Using Microsoft’s Cryptographic API. In Proceedings of the International Conference on Information Security, 2005, pp. 389–401.

An Analysis of Sampling Effect on the Absolute Stability of Discrete-time Bilateral Teleoperation Systems

Amir Aminzadeh Ghavifekr
Faculty of Electrical and Computer
Engineering
University of Tabriz
Tabriz, Iran
aa.ghavifekr@tabrizu.ac.ir

Seyedshahab Chehraghi
Faculty of Electrical and Computer
Engineering
University of Tabriz
Tabriz, Iran
shahab.chehraghi@ieec.org

Giacomo De Rossi
Department of Computer
Science
University of Verona
Verona, Italy
giacomoderossi@outlook.com

Abstract - Absolute stability of discrete-time teleoperation systems can be jeopardized by choosing inappropriate sampling time architecture. A modified structure is presented for the bilateral teleoperation system including continuous-time slave robot, master robot, human operator, and the environment with sampled-data PD-like + dissipation controllers which make the system absolute stable in the presence of the time delay and sampling rates in the communication network. The output position and force signals are quantized with uniform sampling periods. Input-delay approach is used in this paper to convert the sampled-data system to a continuous-time counterpart. The main contribution of this paper is calculating a lower bound on the maximum sampling period as a stability condition. Also, the presented method imposes upper bounds on the damping of robots and notifies the sampling time importance on the transparency and stability of the system. Both simulation and experimental results are performed to show the validity of the proposed conditions and verify the effectiveness of the sampling scheme.

Keywords-Teleoperation System; Sampled-data Control; Stability; Transparency; Networked Control Systems; Master-Slave Robots

I. INTRODUCTION

Teleoperation systems mostly have been utilized in the remote and hazardous operations such as undersea or space explorations [1, 2], and in delicate applications such as micro-assembly and telesurgery [3]. A throughout review of concepts and principles of bilateral teleoperation mechanisms is studied in [4]. Providing stability and transparency in the presence of the unavoidable communication channel time delay is the main challenging topic for researchers in this area. Several continuous-time control approaches such as passivity theorem [5], wave variable method [6], and adaptive controllers [7] have been utilized to address this issue. All well-known robotic theories such as disturbance rejection methods[8, 9], Lyapunov-based controllers[10], and intelligent control [11] have been extended to teleoperation systems.

Although the numerous amount of studies exist for continuous-time bilateral structures, only a few researches have mentioned stability conditions of the discrete-time bilateral structures[12]. One of the most primary challenges in this area is energy leaking due to the using of the zero order hold devices (ZOHs). Numerous methods have been proposed to overcome this issue. These methods including Tustin approach with the scattering theorem [13, 14], the step invariant mapping with appropriate filters [15], input-state stability concept using nonlinear methods [16], and geometric telemanipulation[17]. In [18], it is assumed that the environment is a virtual wall and the stability conditions are calculated for discrete form of this pattern. Extensions for nonidealities such as quantization, friction, and energy losing in [19] are considered as next steps.

A mathematical method for the stability of the discrete-time teleoperation system has been presented in [20, 21]. The passivity and stability of the delay-free discrete-time teleoperators with position-position architectures have been studied in [22] and [23], respectively. In [23], assuming the accurate dynamics of the ZOHs and ideal samplers, the stability of the system is proved using the small gain theorem. In [24] the effect of the sampling rate on the transparency of the teleoperation system has been studied and the hybrid parameters of the discrete-time system have been calculated. Discrete-time circle criterion has been applied to have absolute stable sampled-data haptic interaction [25]. In this method there is no necessity to have passive operator or environment.

This paper evaluates the influence of the sampling rate on the stability conditions of the sampled-data teleoperators. The position-force architecture is selected which means that the transmitting signals are position of the master and force of the slave robots. The stability conditions impose bounds on the damping parameters of the master and slave, and the sampling rate. These analyses prepare mathematical guidelines to design more transparent and stable bilateral teleoperation systems.

The organization of this paper consists of the following sections: Preliminaries and dynamics of the teleoperation systems are introduced in section II. In section III, proposed

discrete-time architecture of bilateral system is presented and the method of finding absolute stability conditions is described. The proposed framework is evaluated for discrete counterpart of the PD-like+dissipation controller in this section to calculate an upper bound for the allowable sampling time without losing the stability. In section IV, the performance of the method has been evaluated by numerical simulations. Finally, an experimental verification is given in section V.

II. TELEOPERATION MODELLING AND DYNAMICS

In the general architecture of teleoperation system, the master robot which is connected to an operator is moved and its position signals are transmitted through the network to the slave side. The standard scheme of this bilateral form is presented in Fig. 1. It is assumed that the dynamics of the robots are similar.

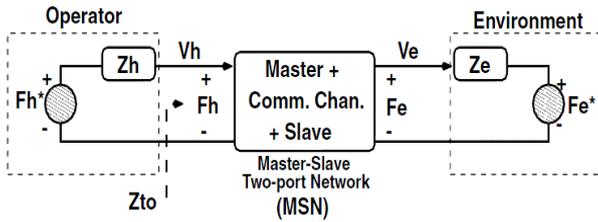


Fig. 1. Standard plot of the bilateral teleoperation system[26]

Assuming master and slave systems as two degree of freedom robots, the well-known dynamics of these systems are written as:

$$m_m \ddot{q}_m + b_m \dot{q}_m = F_h - F_m \quad (1)$$

$$m_s \ddot{q}_s + b_s \dot{q}_s = F_e - F_s$$

where subscripts s and m are implied for the slave and master robots, respectively. The masses and related dampings of robots are given by m and b , respectively. q is position state and \dot{q} is used for velocity signals. F_s and F_m are control torques in the slave and master dynamics. Operator and environment dynamics can be described by:

$$F_h = F_h^* - Z_h(s) s X_m \quad (2)$$

$$F_e = F_e^* - Z_e(s) s X_s \quad (3)$$

where $Z_h(s)$ and $Z_e(s)$ denote The LTI impedances of the human operator and the environment, respectively. F_h^* is the exogenous force input applied by the operator and F_e^* is the exogenous force input applied by the environment. To relate the position and force signals of the master and slave sides the so-called Hybrid matrix can be described:

$$\begin{bmatrix} F_h(s) \\ -sX_s(s) \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} sX_m(s) \\ F_e(s) \end{bmatrix} \quad (4)$$

where

$$h_{11} = Z_m + C_m \frac{Z_s}{Z_s + C_s}, \quad h_{12} = \frac{C_m}{Z_s + C_s} \quad (5)$$

$$h_{21} = -\frac{C_s}{Z_s + C_s}, \quad h_{22} = \frac{1}{Z_s + C_s}$$

where C_s and C_m are local controllers of the slave and master robots, respectively. According to Fig 1, and equations (1-5), it can be deduced that:

$$\frac{X_m}{F_m} = \frac{1}{s m_m s + b_m + z_h}, \quad \frac{X_s}{F_s} = \frac{1}{s m_s s + b_s + z_e} \quad (6)$$

Applying the dynamics of the zero order holds, following transfer functions can be extracted:

$$G_m(s) = \frac{X_m}{F_m^*} = \frac{1}{s m_m s + b_m + z_h} \frac{1 - e^{-sT}}{sT} \quad (7)$$

$$G_s(s) = \frac{X_s}{F_s^*} = \frac{1}{s m_s s + b_s + z_e} \frac{1 - e^{-sT}}{sT}$$

Considering the proposed sampling model and equation (5), controllers can be described as:

$$F_m^*(s) = C_m(e^{sT})[-\alpha G_m^*(s)F_m^*(s) + G_s^*(s)F_s^*(s)] \quad (8)$$

$$F_s^*(s) = C_s(e^{sT})[-G_s^*(s)F_s^*(s) + \alpha G_m^*(s)F_m^*(s)]$$

where $*$ as the superscript indicates the discrete-time counterpart of transform functions. α is a scaling factor related to the position signal. The characteristic equation of the sampled-data teleoperation system can be stated as:

$$1 + \alpha C_m(e^{sT})G_m^*(s) + C_s(e^{sT})G_s^*(s) \quad (9)$$

In [23], using the small gain theorem, the absolute stability of the aforementioned closed loop system is proved. It is declared that the position error-based teleoperation system is stable if and only if satisfies the following inequality:

$$\|M_m N_m + M_s N_s\|_\infty < 1 \quad (10)$$

where

$$N_m = \frac{\alpha b_s C_m(e^{sT})r(s)}{2b_m b_s + \alpha b_s C_m(e^{sT})r(s) + b_m C_s(e^{sT})r(s)}$$

$$N_s = \frac{b_m C_s(e^{sT})r(s)}{2b_m b_s + \alpha b_s C_m(e^{sT})r(s) + b_m C_s(e^{sT})r(s)}$$

$$M_m = -1 + \frac{2b_m}{r(s)} G_m^*(s), \quad M_s = -1 + \frac{2b_s}{r(s)} G_s^*(s) \quad (11)$$

$$r(j\omega) = \frac{T}{2} \frac{e^{-j\omega T} - 1}{1 - \cos \omega T}$$

where T_1 and T_2 are forward and backward delays which are integer multiple of sampling period. Assuming $\alpha = 0$, the aforementioned stability inequality can be described by:

$$\frac{|D + b_s C_m r| + |D + b_m C_s r| + |D|}{|2b_m b_s C_m C_s + b_s C_m^2 C_s r + b_m C_s^2 C_m r + D|} < 1 \quad (12)$$

In which

$$D = \frac{r^2(1 - e^{-(T_1+T_2)s})}{2} \quad (13)$$

III. THE PROPOSED SAMPLED-DATA ARCHITECTURE OF BILATERAL TELEOPERATION SYSTEM

The proposed mathematical structure of the sampled-data teleoperation is presented in this section.

The sampling rate for output signals of both robots assumed to be equal and is denoted by \hat{t}_k , $k \in N$. All sampled position and velocity signals are sent in form of data packets via communication channel which suffers from constant time delay. The samplers in this scheme are time driven while the two zero order holds are event driven. The proposed model is presented in Fig. 2, so that the dash lines are used to illustrate the sampled signals.

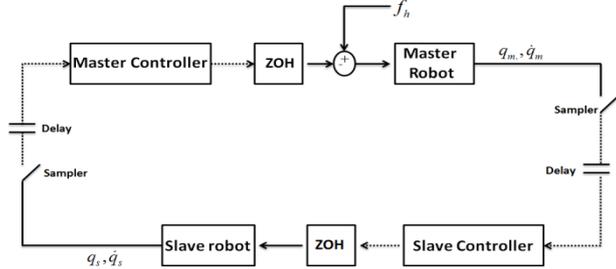


Fig.2. The general structure of the discrete-time teleoperation

It is presumed that the velocity and position signals are sampled at \hat{t}_k . Also, a constant time delay T is applied to state signals during the network transmission. This delay can be larger than interval $[\hat{t}_k, \hat{t}_{k+1}]$. Duration of the k th sampling period is calculated by h_k , i.e. $h_k = \hat{t}_{k+1} - \hat{t}_k$

Assumption 1. There exists $\varepsilon > 0$ such that $\hat{t}_{k+1} - \hat{t}_k > \varepsilon$.

This assumption notifies that sampling intervals cannot perform simultaneously in practical system.

Update rates of the zero order holds at the instants t_k are:

$$t_k = \hat{t}_k + T_k \quad k \in N \quad (14)$$

The duration of last sampling constant \hat{t}_k is calculated as:

$$\mu(t) \triangleq t - \hat{t}_k = t - t_k + T \quad (15)$$

$\mu(t)$ is defined as the induced delay. The maximum amount of this network-induced delay represented by γ is calculated as:

$$\gamma = \sup(\mu(t)) = \sup(\hat{t}_{k+1} - \hat{t}_k) \quad (16)$$

The utilized controllers for master and slave robots in the proposed structure of Fig. 2 are PD like controllers + dissipations. An emulated mode of this controller is proposed in [27] to improve the accuracy of the force tracking and achieve better coordination of robots. The continuous-time forms of controllers are proposed as:

$$\begin{aligned} \tau_1(t) &= -K_v(\dot{q}_1(t) - \dot{q}_2(t - \tau_2)) - (K_d + P_\varepsilon)\dot{q}_1(t) - K_p(q_1(t) - q_2(t - \tau_2)) \\ \tau_2(t) &= -K_v(\dot{q}_2(t) - \dot{q}_1(t - \tau_1)) - (K_d + P_\varepsilon)\dot{q}_2(t) - K_p(q_2(t) - q_1(t - \tau_1)) \end{aligned} \quad (17)$$

where $T_1, T_2 \geq 0$ are delays from master to slave and vice versa. K_v, K_p are the symmetric and positive definite gains, K_d is the positive dissipation gain and P_ε is an extra damping to protect master-slave coordination. It is proved

in [27], that choosing $K_d = \frac{\nu}{2}K_p$ where $\nu > 0$ is an upper bound of the general delay $T_1 + T_2$ leads to have a passive teleoperation system. Also, if the operator and the environment are passive, the position tracking error between robots will be bounded. Furthermore, if the velocity and accelerations signals converge to the zero, force tracking error will be achieved.

The primary discrete-time form of control signals in (17) can be rewritten as:

$$\begin{aligned} F_m(t) &= -K_v(\dot{x}_m(\hat{t}_k) - \dot{x}_s(\hat{t}_k - T_2)) - (K_d + P_\varepsilon)\dot{x}_m(\hat{t}_k) - K_p(x_m(\hat{t}_k) - x_s(\hat{t}_k - T_2)) \\ F_s(t) &= -K_v(\dot{x}_s(\hat{t}_k) - \dot{x}_m(\hat{t}_k - T_1)) - (K_d + P_\varepsilon)\dot{x}_s(\hat{t}_k) - K_p(x_s(\hat{t}_k) - x_m(\hat{t}_k - T_1)) \end{aligned} \quad (18)$$

It is notable that the gains of controllers for slave and master robots are equal. This is due to the similar dynamics of these robots.

The input-delay method is proposed in [28] to calculate the maximum allowable network delay which help to preserve the exponential stability of the discrete-time systems. Using this

Approach, (18) can be rewritten as:

$$\begin{aligned} F_m(t) &= -K_v(\dot{x}_m(t - \mu_s) - \dot{x}_s(t - \mu_s - T_2)) - (K_d + P_\varepsilon)\dot{x}_m(t - \mu_s) - K_p(x_m(t - \mu_s) - x_s(t - \mu_s - T_2)) \\ F_s(t) &= -K_v(\dot{x}_s(t - \mu_s) - \dot{x}_m(t - \mu_s - T_1)) - (K_d + P_\varepsilon)\dot{x}_s(t - \mu_s) - K_p(x_s(t - \mu_s) - x_m(t - \mu_s - T_1)) \end{aligned} \quad (19)$$

By substituting the controllers (19) in the absolute stability condition of (11) and using bilinear transformation method we have:

$$\tau_1(t) = \tau_2(t) = -K_v \frac{z-1}{Tz} - (K_d + P_\varepsilon) \frac{z-1}{Tz} - K_p \quad (20)$$

Thus, the stability condition can be simplified to:

$$b_m, b_s > K_p T + 2K_d - 2P_\varepsilon - 2K_v \quad (21)$$

IV. NUMERICAL SIMULATION RESULTS

The proposed stability conditions have been tested on the 1-DOF teleoperation system modeled by the mass and damping terms. By simulation, the effect of sampling rate on the behavior of sampled-data teleoperation system has been studied. The gains for the PD-like with dissipation controller are chosen $K_p = 1, K_d = 2, P_\varepsilon = 0.002, K_v = 10$.

The environment acts like a stiff wall, which reflecting the overall torque of the slave robot. The following scenario has been considered in simulation. A step force has been applied to the master robot by the human operator for 10 seconds from 10s till 20s. The slave robot meets the environment at $4rad$. The spring-mass structure is used to model the human operator. Spring coefficient is $10N/m$ and damping gain is chosen $1Ns/m$. The generated force by the operator is illustrated in Fig.3

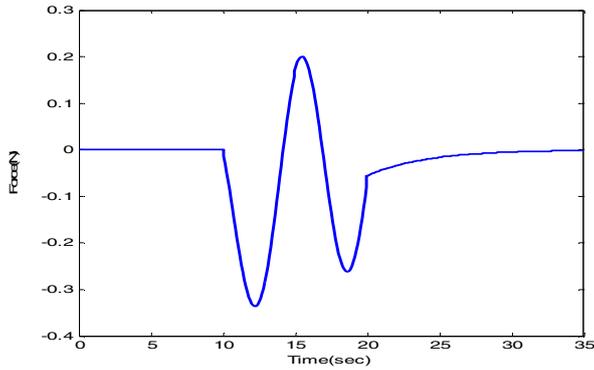


Fig. 3. External force generated by the operator

Both master and slave systems are assumed as one degree of freedom robots with transform function of $M(s) = 2 / (2 + s)$. Fig. 4 and Fig. 5 present the position and force signals of the master and slave robots, respectively. The maximum allowed sampling period is chosen 0.006s according to (21).

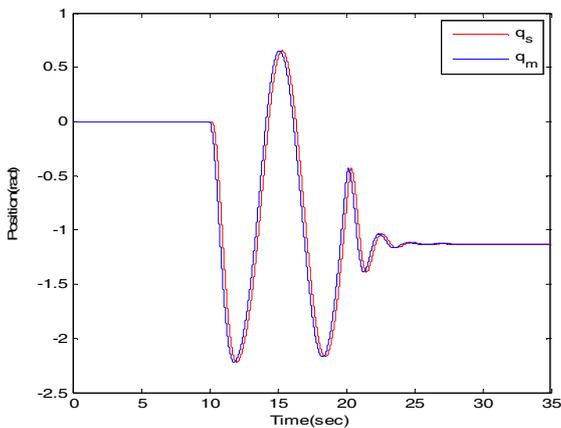


Fig. 4. Master and slave position signals for sampled-data counterpart of PD-like+dissipation controller

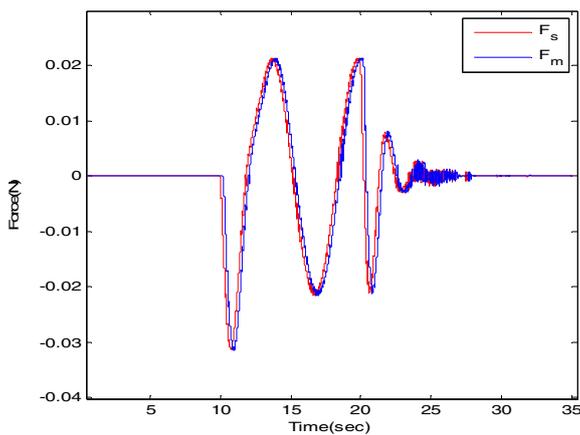


Fig. 5. Master and slave force signals for sampled-data counterpart of PD-like+dissipation controller

The most noticeable remark is the stability can be jeopardized by increasing gains of controllers and sampling time. Also, the physical and practical parameters of the robot cannot be varied frequently. Thus, there should be a trade-off between sampling rate and controller characteristics.

V. EXPERIMENTAL SETUP AND RESULTS

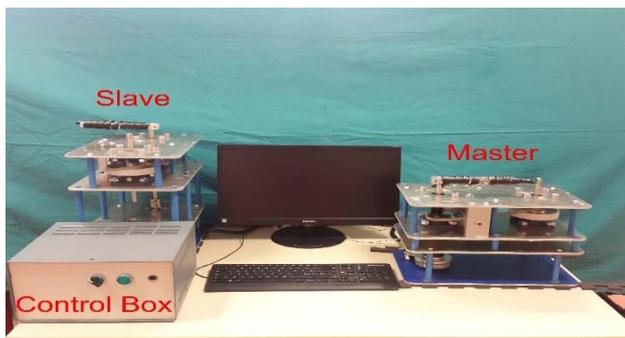
The system has been tested on a physical setup to establish the validity of the simulation's results. The setup is two brushless DC motors attached to two gearboxes for torque multiplying and two handles. One degree of freedom force sensor is attached to each handle to measure the amount of force/torque of the user and the environment exert on the system. The motor actuators are connected via EtherCAT to a PC, running a soft real-time Linux kernel. This platform provides high frequency software control and allows testing different teleoperation configurations, such as position-position (P-P), force-position (F-P), and 4-channel architecture. By using native kernel libraries and C++ programming, it is possible to guarantee high performances given any desired controller for the system.

Working on a physical setup introduces limitations in both the controller output and the measurement obtainable by the sensors that are usually neglected in a simulation. For instance, the motors accept voltage control in the range of $\pm 5V$ and both position and force sensors produce quantized and noisy signals that require filtering. Additionally, the register used to store the encoder value overflows after approximately fifty full rotations. Therefore, testing a system in simulation and physical setup can be instructive to see whether the sampling time condition or the passivity conditions still hold.

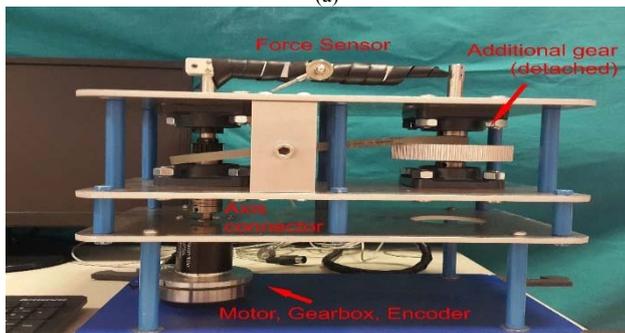
The simulation has been enriched by introducing the parameters resulting from the setup identification process. The latter involved the extraction of data from the motor datasheet and the linear regression over acquired samples from free-runs of the system. These parameters are: 50 Hz for the first-order velocity filters cutoff frequency, according to the spectral analysis of the raw data; 4.054 V/N force-to-voltage coefficient (from motor and force sensor datasheet comparison); 0.1 m arm length for the lever attached to the motor; $2\pi / 4096$ rad/step quantization for the step encoder (from datasheet). The noise of the sensor has been represented as white Gaussian noise ($0\mu, 1\sigma$). Finally, the motor, with the attached gearbox and lever, has been identified by using a Kalman smoother on the velocity signal. The result is the following first-order continuous-time transfer function:

$$M(s) = 19.34 / (1.217s + 1), \quad \text{with rotor inertia } J = 23.54 \text{ kgm}^2 \text{ and viscous friction coefficient } F = 0.0517.$$

The general scheme of the system is depicted in Figure 6. The first part represents the entire system including master and slave, while the second part shows the details of the master manipulator.



(a)



(b)

Fig. 6. a) General scheme of the experimental system b) Components of the master manipulator

The configuration used in this experiment is a stable one with $K_p = 8.4$ and $K_d = 0.0005$. The environment is placed approximately at 3.5rad , as in the simulation. The evaluated maximum sampling time for stability, obtained by the evaluation of equation (21), is $T = 0.006\text{s}$; due to internal time delays in the control loop, such value decreases down to $T = 0.003\text{s}$. When this limit is reached, the system presents an unstable behavior.

The position and force tracking signals, presented in Figures 7 and 8, show a scenario in which the slave suddenly hits the obstacle with three different velocities. In the first two contacts, the controllers are quick to stabilize the motion; in the third contact, the greater impact force carried by the increased momentum results in a finite oscillation, which finally stops near 2.5rad .

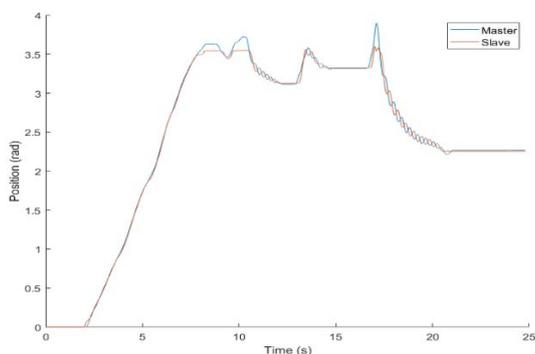


Fig. 7. Master and slave robots position signals for experimental discrete-time PD-like+ dissipation controller

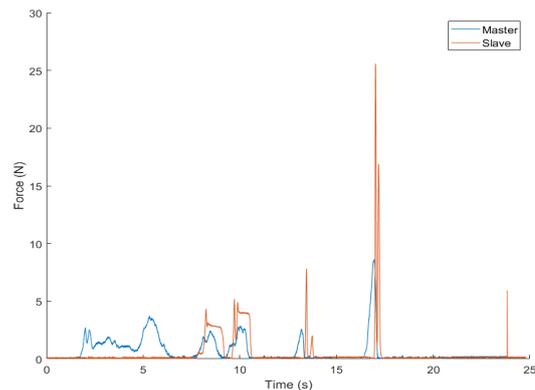


Fig. 8. Master and slave robots force signals for experimental discrete-time PD-like+ dissipation controller

VI. CONCLUSION

Despite extensive researches about continuous-time teleoperation systems, the sampled-data structures have not been studied widely in the control literature. The stability condition of these systems depends on the rate of the sampling time and unavoidable delay of the communication channel. In this paper, according to the proposed method, limitations of the passivity conditions are omitted due to the non-passive equations of slave and master dynamics and arbitrary passive models of the operator and environment. It is noticeable that although selecting larger controller gains can provide better transparency in the continuous-time structures, there should be a trade-off between the stability conditions and transparency of the system in the discrete-time structures. As a suggestion for future studies on the discrete-time teleoperators, the variable time delay of the network and the quantization error effects of the samplers can be taken into account. Also, the proposed stability conditions can be extended to the 4-channel architecture of the bilateral teleoperation systems.

References

- [1] E. F. Cardenas and M. S. Dutra, "An Augmented Reality Application to Assist Teleoperation of Underwater Manipulators," *IEEE Latin America Transactions*, vol. 14, pp. 863-869, 2016.
- [2] K. Némethy, J. Gáti, G. Kártyás, and F. Hegyesi, "Exoskeleton and the remote teleoperation projects," in *Applied Machine Intelligence and Informatics (SAMI), 2018 IEEE 16th World Symposium on*, 2018, pp. 000073-000080.
- [3] S. Ryu and G.-H. Yang, "Telesurgery system using surgical master device type of 3PUU," in *Ubiquitous Robots and Ambient Intelligence (URAI), 2017 14th International Conference on*, 2017, pp. 546-549.
- [4] P. F. Hokayem and M. W. Spong, "Bilateral teleoperation: An historical survey," *Automatica*, vol. 42, pp. 2035-2057, 2006.

- [5] H.-C. Hu and Y.-C. Liu, "Passivity-based control framework for task-space bilateral teleoperation with parametric uncertainty over unreliable networks," *ISA transactions*, 2017.
- [6] D. Sun, F. Naghdy, and H. Du, "Application of wave-variable control to bilateral teleoperation systems: A survey," *Annual Reviews in Control*, vol. 38, pp. 12-31, 2014.
- [7] Z. Chen, Y.-J. Pan, and J. Gu, "Adaptive robust control of bilateral teleoperation systems with unmeasurable environmental force and arbitrary time delays," *IET Control Theory & Applications*, vol. 8, pp. 1456-1464, 2014.
- [8] A. A. Ghavifekr, M. Badamchizadeh, G. Alizadeh, and A. Arjmandi, "Designing inverse dynamic controller with integral action for motion planning of surgical robot in the presence of bounded disturbances," in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, 2013, pp. 1-6.
- [9] A.-A. Ghavifekr, "Evaluation of Three Nonlinear Control Methods to Reject the Constant Bounded Disturbance for Robotic Manipulators," *Majlesi Journal of Mechatronic Systems*, vol. 1, 2012.
- [10] A. Ghiasi, A. Ghavifekr, Y. S. Hagh, and H. SeyedGholami, "Designing adaptive robust extended Kalman filter based on Lyapunov-based controller for robotics manipulators," in *2015 6th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO)*, 2015, pp. 1-6.
- [11] A. Ghavifekr, S. Ghaemi, and R. Behinfaraz, "A Modified biogeography based optimization (bbo) algorithm for time optimal motion planning of 5 dof pc-based gryphon robot."
- [12] A. A. Ghavifekr, A. R. Ghiasi, and M. A. Badamchizadeh, "Discrete-time control of bilateral teleoperation systems: a review," *Robotica*, vol. 36, pp. 552-569, 2018.
- [13] R. J. Anderson, "Building a modular robot control system using passivity and scattering theory," in *Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on*, 1996, pp. 698-705.
- [14] K. Kosuge and H. Murayama, "Bilateral feedback control of telemanipulator via computer network in discrete time domain," in *Robotics and Automation, 1997. Proceedings., 1997 IEEE International Conference on*, 1997, pp. 2219-2224.
- [15] G. Leung and B. Francis, "Bilateral control of teleoperators with time delay through a digital communication channel," in *Proceedings of the annual allerton conference on communication control and computing*, 1992, pp. 692-692.
- [16] I. Polushin and H. Marquez, "Stabilization of bilaterally controlled teleoperators with communication delay: an ISS approach," *International Journal of Control*, vol. 76, pp. 858-870, 2003.
- [17] S. Stramigioli, "About the use of port concepts for passive geometric telemanipulation with varying time delays," in *Proceedings to Mechatronics Conference*, 2002.
- [18] J. J. Gil, A. Avello, A. Rubio, and J. Florez, "Stability analysis of a 1 dof haptic interface using the routh-hurwitz criterion," *Control Systems Technology, IEEE Transactions on*, vol. 12, pp. 583-588, 2004.
- [19] N. Diolaiti, G. Niemeyer, F. Barbag, and J. K. Salisbury Jr, "Stability of haptic rendering: Discretization, quantization, time delay, and coulomb effects," *Robotics, IEEE Transactions on*, vol. 22, pp. 256-268, 2006.
- [20] M. Tavakoli, A. Aziminejad, R. Patel, and M. Moallem, "Discrete-time bilateral teleoperation: modelling and stability analysis," *IET Control Theory & Applications*, vol. 2, pp. 496-512, 2008.
- [21] A. A. Ghavifekr, A. R. Ghiasi, M. A. Badamchizadeh, F. Hashemzadeh, and P. Fiorini, "Stability analysis of the linear discrete teleoperation systems with stochastic sampling and data dropout," *European Journal of Control*, vol. 41, pp. 63-71, 2018.
- [22] A. Jazayeri and M. Tavakoli, "A passivity criterion for sampled-data bilateral teleoperation systems," *Haptics, IEEE Transactions on*, vol. 6, pp. 363-369, 2013.
- [23] A. Jazayeri and M. Tavakoli, "Absolute stability analysis of sampled-data scaled bilateral teleoperation systems," *Control Engineering Practice*, vol. 21, pp. 1053-1064, 2013.
- [24] T. Yang, Y. L. Fu, and M. Tavakoi, "An analysis of sampling effect on bilateral teleoperation system transparency," in *Control Conference (CCC), 2015 34th Chinese*, 2015, pp. 5896-5900.
- [25] N. Miandashti and M. Tavakoli, "Stability of sampled-data, delayed haptic interaction under passive or active operator," *IET Control Theory & Applications*, vol. 8, pp. 1769-1780, 2014.
- [26] K. Hashtrudi-Zaad and S. E. Salcudean, "Analysis of control architectures for teleoperation systems with impedance/admittance master and slave manipulators," *The International Journal of Robotics Research*, vol. 20, pp. 419-445, 2001.
- [27] D. Lee and M. W. Spong, "Passive bilateral teleoperation with constant time delay," *IEEE transactions on robotics*, vol. 22, pp. 269-281, 2006.
- [28] E. Fridman, "A refined input delay approach to sampled-data control," *Automatica*, vol. 46, pp. 421-427, 2010.

The Software Platform for Evaluation of Effectiveness of Network Systems Analysis Technologies

Olha Ponomarenko
Department of Electronic
Computers
Kharkiv National University of
Radio Electronics
Kharkiv, Ukraine
olha.ponomarenko@nure.ua

Valeriy Gorbachov
Department of Electronic
Computers
Kharkiv National University of
Radio Electronics
Kharkiv, Ukraine
valeriy.gorbachov@nure.ua

Abdulrahman Kataeba Batiaa
Department of Electronic
Computers
Kharkiv National University of
Radio Electronics
Kharkiv, Ukraine
kotaeba04@gmail.com

Oksana Kotkova
Department of Electronic
Computers
Kharkiv National University of
Radio Electronics
Kharkiv, Ukraine
oksana.kotkova@nure.ua

Abstract—Networks are used as a common model of a wide variety of complex systems, including social, biological, information, and technological domains. Nodes represent components of the system and links indicate interactions between them. Network monitoring, fast decision making and modeling techniques are fundamental to topology research of network systems. Topological analysis of networks is needed to develop network planning and network management, bottleneck and failure detection algorithms, system performance evaluation. The main objective of this paper is to research on network topology and use a software platform to evaluate the effectiveness of topology formal transformations for reducing the system dimension. The platform includes the set of modules: evaluation of topological network parameters, equivalent topological transformations, maximum flow searching, and topology generator. The experiment allowed to evaluate the effectiveness of both the network topology formal transformations and the effectiveness of the platform itself.

Keywords—large scale system, topological transformations, software platform, topology generator, maximum flow problem, UML diagram

I. INTRODUCTION

Over the past decade, representations and studies of complex systems have been associated with so-called complex networks, which are network-based representations of complex systems. Complex systems are systems in which the pattern of interactions between a system's constituent parts is itself complex and is evolving together with the system's dynamics. In the context of network theory, a complex network is a network (graph) with non-trivial topological features and with a multitude of non-trivial statistical challenges [1], [2].

It is not uncommon now to see networks with millions or even billions of vertices. An increase in the size of networks leads to the development of new analytical approaches for their presentation and performance evaluation. When developing such approaches, researchers face various kinds of problems.

For networks consisting of even several dozens of vertices, it is quite simple to draw a picture of the network and answer

specific questions about the structure of the network by studying this picture. This has been one of the primary ways to gain an understanding of network structure. Nowadays, a variety of great visualization tools are available, which helps to structure and to visualize the networks [3]. However, they are useless for an analysis of networks consisting of a million or a billion vertices.

In recent years, complex network theory becomes more and more popular. This theory is based on a solid mathematical framework that aims to solve a range of complex problems.

1. Development of an appropriate structural description of a complex system to determine the elements, subsystems and connections among them.

2. Development of approaches of determination and prediction of statistical properties, that characterize the structure and behavior of networked systems. For example, the definition of strongly connected components, shortest paths, cycles, races, etc. In addition, the system structure model is used to analyze the quality metrics of the structure. For example, clustering problems, network correlations.

3. Implementation of aggregation and decomposition technologies to reduce the dimension of the system, when the time of performance evaluation plays an important role.

4. Optimal structural design. Most of problems of this group are the problems of increased complexity, such as, evaluation of the effects of structure on system behavior, equivalent transformations of the topological structure reducing the dimension of system, redistribution of links of the established structure, bottlenecks detecting.

Network systems require specific methods for analysis and design. The main objective of this paper is to research on network topology and use a software platform to evaluate the effectiveness of various systems analysis technologies including network dimensions reduction technology. The software platform, considered in the paper, includes the set of modules: evaluation of topological network parameters, topology

generator, equivalent topological transformations, and maximum flow searching.

II. THE SOFTWARE PLATFORM

The module of topology generator. Network researchers often need to perform the preliminary evaluation of new designs from point of view of effectiveness of network topology, network capability to withstand high loads and remain operational. Due to the immense scale of modern network systems, creation a real system for the purpose of experimental study is nearly impossible. In this case, researchers evaluate proposed solutions using generated networks. In the paper, a generator is used to evaluate the effectiveness of the software platform.

There are a wide variety of generators available to the research community. Some of them mainly aim to generate random topologies [4], others aim to imitate the hierarchical properties of the Internet [5], [6] and still others aim to reproduce degree-related properties of the Internet [7], [8]. Each of these generators implement a different set of generation models. An overview of generators shows that a unified model that considers both hierarchical properties, degree distribution properties, connectivity properties and incorporate casual models has not yet been developed. However, some of the requirements for a network topology generator, listed by [9], include the following.

Representativeness: The generated topologies must be accurate, based on the input arguments such as hierarchical structure and degree distribution characteristics.

Flexibility: In the absence of a universally accepted model, the generator should include different methods and models.

Extensibility: The tool should allow the user to extend the generator's capabilities by adding their own new generation models.

Efficiency: The tool should be efficient for generating large topologies while keeping the required statistical characteristics intact. This can make it possible to test real world scenarios.

In the paper the degree distribution-based generator has been implemented. This type of generators more accurately captures the large-scale structure of studied topologies [10].

The module of equivalent topological transformations. The main purpose of the module is to find an equivalent simpler representation of network systems while preserving the characteristic properties of the higher dimension system. The module is based on the approach related to formal transformations of the system structure model. The multilevel aggregation has been applied to obtain a reduction in computational complexity and faster modeling [11]. The approach uses as input the matrix form of the system topology representation. As output, the approach yields the matrix form of simplified structure of the system.

The module of maximum flow searching. Depending on the problem being solved, this module can solve such tasks: network designing and network management; detection of bottleneck, deadlocks and failures; maximum flow searching; system performance evaluation. In the work, in the module the problem

of maximum flow is implemented [12]. The maximum flow problem belongs to the group of topological analysis problems. Its purpose is to distribute network flows to achieve the maximum values of communication efficiency. The maximum flow problem is formulated as follows: the maximum possible total value of the flow between the source and the sink has to be found for given network with established initial distribution of flows for graph edges and capacities. It means that the flow has to be increased if it has not reached the maximum value. The maximum flow value is equal to the sum of weights of the edges in the minimum cut in accordance with the theorem proved by Ford and Fulkerson, which is applied for solving the maximum flow problem [13].

III. IMPLEMENTATION OF THE SOFTWARE PLATFORM

The generator of structural models is implemented in the Java programming language (Fig. 1).

The Generator class was created, which contains fields of the type GraphStructure and SystemStructure. The GraphStructure class represents a graph and contains the vertices of the graph which are shown by the Vertex class and the edges which are shown by the Edge class. The Vertex class contains the vertex number and its degree. The Edge class contains numbers of vertices which associated by the edge. Also, the class contains the matrix, in which the edges generation of the graph is performed.

The system structure is represented by the SystemStructure class. It contains elements of the system which are shown by the Element class and connections between the elements which are shown by the Connection class. The Element class contains the element number and the amount of input and output contacts. The Connection class contains the numbers of the element and the input contact, in which a connection enters, the numbers of the element and the output contact, from which the connection exits, and channel capacity.

At the beginning of the algorithm the graph generation is fulfilled. First of all, an ArrayList collection of Vertex objects are created. After that, the edges generation of the graph is performed and the data is entered into the matrix. Following that, an ArrayList collection of Edges objects is created.

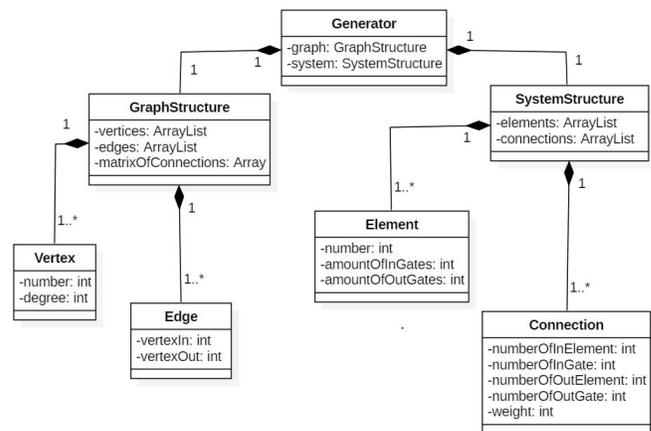


Fig. 1. Simplified UML diagram with class names and fields

At the next step, the system structure is created on the basis of the graph which was obtained. The edges of the graph are traversed and the ArrayList collections of the Element and Connection objects are created. The collection of Connection objects contains all the connections of the elements in the system. It is a crucial item, because the table of the elements connections of the system is constructed based on the collection. Finally, the table is transmitted to the block of the composition method.

The Ford-Fulkerson algorithm is implemented in the C# programming language.

The program model is represented as a UML class diagram in Fig. 2.

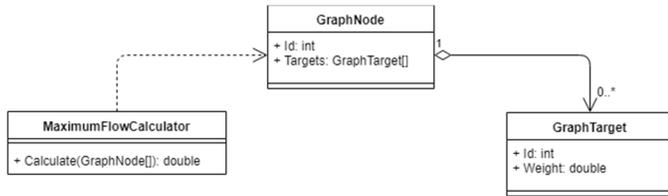


Fig. 2. UML diagram of the program model

In the software implementation, the additional classes Program and FileGraphNodesProvider is used.

To handle input files, a FileGraphNodesProvider class has been created. It deserializes data and creates objects of the GraphNode and GraphTarget classes, which are designed to store system nodes, their interconnections and throughput capabilities.

To calculate the maximum flow in the system by the Ford-Fulkerson theorem, the MaximumFlowCalculator class is created.

The Program class is the entry point to the program. It is designed to process command line parameters, call methods of the FileGraphNodesProvider and MaximumFlowCalculator classes, and display the results of the application.

IV. EVALUATION OF THE SOFTWARE PLATFORM EFFECTIVENESS

The main objective of the experiment is to evaluate the effectiveness of both the network topology formal transformations and the effectiveness of the platform itself. The platform includes the set of modules: evaluation of topological network parameters, topology generator, equivalent topological transformations, and maximum flow searching. The procedure of experiment consists in the following.

The topology generator module generates network systems. These systems generated by the generator have topological characteristics similar to those of a network system with a number of elements equal to 12. At the next step, the equivalent topological transformations module performs three-level topological transformations. At the last step, maximum flow searching module solves the Ford and Fulkerson problem.

Analysis of experimental outcomes. The runtime of the modules of topology generator and maximum flow searching (Fig. 3) is significantly less than the total runtime of the problem

solve (Fig. 4). This runtime practically coincides with the runtime of the equivalent topological transformations module.

The analysis of the graphs of the runtime of modules leads to the conclusion that with the increase in the number of elements and the links among them, the runtime of the modules rises steeply. Perform an analysis of the results for each module.

The algorithm of the network systems generator, at the time of the formation of a new element, has to go through the collection that stores the arcs of the graph, and also check whether such element already exists. To do this, it needs to analyze all elements of the system. If the number of system elements increases, the runtime of generator also increases.

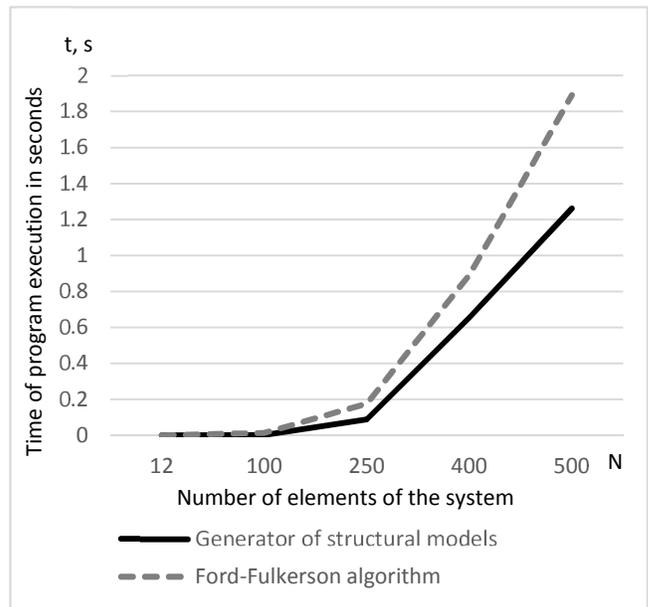


Fig. 3. The runtime of generator and Ford-Fulkerson algorithms

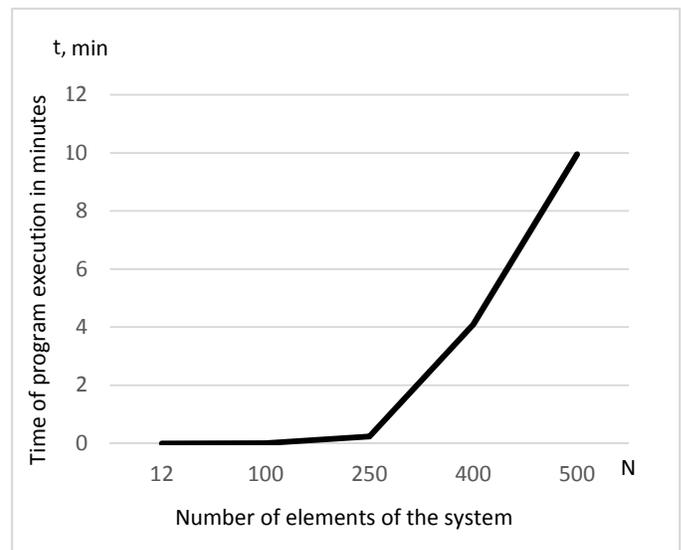


Fig. 4. The total runtime

The algorithm of the multi-level transformation is based on combining elements of the previous level into subsystems and forming links within these subsystems and among subsystems. The increase in the number of system elements leads to a significant increase in the dimension of the matrix forms that represent the structure of the network system. The increase in the dimension of the matrix forms, in turn, leads to an increase in the runtime of algorithm at the stage of forming fictitious contacts of the subsystems. The time of the formation of links among subsystems also increases, because of analysis of enormous number of contacts of the subsystems. Taking into account the fact that high-dimensional matrices are dispersed, a lot of time is spent on unproductive operations.

With an increase in the number of system elements, the runtime of the maximum flow searching algorithm extremely increases. This is due to the fact that when calculating the maximum flow, the list of elements, their links and link capacities are converted from a list into a matrix of connections. Operating with such matrix makes the maximum flow searching algorithm time consuming.

V. CONCLUSION AND FUTURE WORK

The main conclusion related to the software platform consists in the following. The platform aims to do four things. First, to find statistical properties which characterize the structure of networked systems and create generative models. Second, to reduce the dimension of network systems. Third, to evaluate how will network structure affect on the system performance. Fourth, to design optimal network system.

An increase in the size of network systems leads to the essential rise of the general runtime of computation. The main reason is the matrix forms that represent the structure of network system at all stages of problem solving. This is an important conclusion in understanding the direction of future research. Future studies should aim to develop an effective structural description of network systems.

REFERENCES

- [1] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*, Springer, New York, 2009.
- [2] M. Newman, *Networks: An Introduction*, Oxford University Press, 2010.
- [3] J. Hackl, "tikz-network: a LaTeX library for visualizing complex networks," in 6th International Conference on Complex Networks & Their Applications, Lyon, France, 2017.
- [4] B. Waxman, "Routing of multipoint connections," in *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, 1988, pp. 1617–1622.
- [5] K. Calvert, M. Doar and E. Zegura, "Modeling Internet topology," *IEEE Communications Magazine*, vol. 35, no. 6, 1997, pp. 160–163.
- [6] M. Doar, "A Better Model for Generating Test Networks," in *Global Telecommunications Conference GLOBECOM*. IEEE, 1996, pp. 86–93.
- [7] S. Jamin and J. Winick, "Inet-3.0: Internet Topology Generator," University of Michigan, Ann Arbor, 2002.
- [8] A. Medina, I. Matta and J. Byers, "On the origin of power laws in Internet topologies," *ACM SIGCOMM Computer Communication Review*, vol. 30, no. 2, 2000, pp. 160–163.
- [9] A. Medina, A. Lakhina, I. Matta and J. Byers, "BRITE: An approach to universal topology generation," in *Proceedings of the Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, Cincinnati, OH, USA, 15–18 August 2001, pp. 346–353.
- [10] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker and Walter Willinger, "Network Topology Generators: Degree-Based vs. Structural," in *SIGCOMM'02 Proceedings of the Conference on Applications, Technologies, Architectures and Protocols for Computer Communications*, 2002, pp. 147–159.
- [11] V. Gorbachov, A. K. Batiaa, O. Ponomarenko and Y. Romanenkov, "Formal transformations of structural models of complex network systems," in *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies DESSERT'2018*, Conference proceedings, Kyiv, 2018, pp. 473–477.
- [12] T. Cormen, C. Leiserson, R. Rivest and C. Stein, *Introduction to algorithms*, Third Edition, The MIT Press, 2009.
- [13] L. Ford and D. Fulkerson, "Maximal Flow Through a Network," *Canadian Journal of Mathematics*, vol. 8, 1956, pp. 399–404.

Multidimensional Hierarchical Model of Behavioral Testing of Distributed Information Systems

Oleksandr Martynyuk
*Department of Computer Intelligent
Systems and Networks*
*Odessa National Polytechnic
University*
Odessa, Ukraine
anmartynyuk@ukr.net

Oleksandr Drozd
*Department of Computer Intelligent
Systems and Networks*
*Odessa National Polytechnic
University*
Odessa, Ukraine
drozd@ukr.net

Hanna Suhak
*Department of Computerized Control
Systems*
*Odessa National Polytechnic
University*
Odessa, Ukraine
hanna.suhak@gmail.com

Dmitry Martynyuk
*Department of Computer Intelligent
Systems and Networks*
Odessa National Polytechnic University
Odessa, Ukraine
d.hnr@gmail.com

Lyudmila Sugak
Educational Methodical Department
Odessa National Polytechnic University
Odessa, Ukraine
lpsugak@ukr.net

Abstract — Perspective distributed information systems (DIS), along with the often present property of criticality of use, increasingly possess the agent properties of autonomy, mobility, intelligence, cooperativeness. Automated systems of technical diagnostics (ASTD) of such DIS in accordance with the first property should ensure their complete real-time testing, and for both static hardware-software and information environment of DIS placement, and for a dynamic set of information-control flows corresponding to the main functions and scenarios of her work, that is, her behavior. DIS agent properties only increase the need and deepen the content of testing this behavioral DIS model by ASTD. This paper discusses a complex model of behavioral check for DIS components, in particular, for cooperativeness of agents, based on a multidimensional, multi-level, heterogeneous and multi-purpose structure of their behavioral interactions. In accordance with this, both within each level, and between the levels of the near circle (adjacent) and the far circle (connected by one or more intermediate levels), behavior check and recognition are performed for input models of the automaton class. The check model has features, firstly, determining the structure of identifiers, atomic and compositional checks, corresponding to the multidimensional and multilevel structure of behavioral interactions; secondly, the signatures of special operations and relations of cooperation for these checks; thirdly, the basic laws of reception/transfer, transformation, inheritance, encapsulation, preservation/accumulation. The behavioral check model allows defining the basic conditions for constructing multi-level DIS verification methods, reducing the time and resource costs due to decomposition of the check analysis, in particular, multi-agent hierarchical, which is important for real-time verification.

Keywords — *Distributed Information Systems, Behavioral Testing, Check Model, Identifier, Check Fragment*

I. INTRODUCTION

Perspective distributed information systems (DIS) [1-3], rapidly expanding their scope of application [4], are characterized by a sharp complication of the problems solved with their help, with a significant increase in the criticality of their application, growing intelligence with the use of fuzzy, evolutionary and neural methods, and the speed bordering on the real time of the functioning of objects of the domain [5-9]. Component autonomy and mobility, as well as the growth of the degree and structure of their interactions, both internal -

intercomponent and external - with external objects of the domain and the global network [10, 11], are becoming more and more significant, developing features of modern DIS. These properties of DIS acquire the character of general and indicate that they obtained the properties of multi-agent and dynamic [17] systems that promptly form special distributed and shared structures of tasks, resources and processes in the environments of the accommodation infrastructure. It should be noted that a significantly higher level of dynamic, situational communications, coordination and cooperation in such systems exacerbates the risks of access, uncertainty, functional disability, failures and errors, incorrect and malicious actions [18-22]. As a rule, these risks are reduced or even eliminated by a set of security measures and information protection systems, for example, by means of authorization/authentication, digital signature, encryption, attributes and access rights/trust, multilevel screening, subject-logical virtualization [23-26]. However, a higher level of reliability of DIS functioning is provided by additional means of their formal check and diagnosis [27-32]. Thus, analysis, design, maintenance of DIS, often NP-complex [33-36], with the use of many complex technologies [37-40], with a sharp increase in the degree of efficiency of both their construction and verification, in particular, based on behavioral online and offline testing and diagnosis [36-38], used special FPGA testing [37, 39-41].

For the analysis of component cooperations and DIS, in general, commonly used methods are deterministic, probabilistic, fuzzy, evolutionary check, diagnosis and testing of their structural, functional and informational properties and mechanisms, characterized by acceptable values of the reliability of work and resource costs. At the same time, the significantly increased dynamism, uncertainty, intelligence and diversity of situational component DIS cooperations, both in the distribution within the fuzzy boundaries and in the tasks to be solved, impose ever more stringent time requirements for the existing methods of check, diagnosis and testing up to real of time. This limits the use of most of the known methods to systems of medium complexity.

Solving these problems uses hardware methods, decomposition and paralleling of the presented check and diagnostic analysis, and can also be based on the development

of formal models, expansion of the class of analyzed properties and mechanisms of DIS, in particular, additional research of testing of multidimensional, hierarchical, heterogeneous and multi-purpose behavioral models on base of systems of hierarchical Petri nets (PN) - for situational spatial-temporal cooperations of components and processes into DIS.

The relevance of the work is due to the need to develop the existing methods of decomposition, behavioral testing of DIS with the features of situational dynamic multidimensional, hierarchical spatial-temporal, multiagent cooperations with the accumulation of knowledge of check behavior, in particular, the near (frequent, adjacent near) and distant (episodic, indirectly remote) circle, represented by systems of hierarchical PN. The proposed technology of behavioral testing of DIS increases the efficiency of the formation of verified, secure, dynamic, spatial-temporal, distributed systems of tasks, resources and processes for DIS. The result is a reduction in the time of analysis, testing and recovery of DIS, close to real time.

II. BUILDING THE MULTIDIMENSIONAL HIERARCHICAL MODELS OF BEHAVIOURAL TESTING OF DIS

A formal model of behavioral control determines the conditions for its implementation, taking into account which one or another method of behavioral control is built. The DIS models built, for example, on the basis of PN, simple and extended, in particular hierarchical, have great flexibility, power, and expressiveness. Representing many asynchronous parallel processes using the chip mechanism, simple and extended Petri nets allow you to implicitly and naturally implement spatial, time-parallel decomposition of the DIS model, hierarchical PN additionally provide the possibility of its temporal, time-consistent decomposition. The extended PN $S(f)$ can have a fairly general form, in particular:

$$S(f)=(P, T, X, Y, In, Pb, Ep, Et, F, S, M_0, L, K), \quad (1)$$

where P, T are the sets of positions and transitions, X, Y are the sets of input and output signals, respectively, for variable conditions, events, actions and functions, arising and placed in positions and transitions; $In \subset \mathbb{N}$ $In \square \mathbb{N}$ is the set of integer time intervals of transitions; $Pb \subset [0; 1] \subset \mathbb{D}$ is the set of probability coefficients in the range $[0; 1]$; $Ep \subset \mathbb{N}$ - a set of integer energy costs for the formation of conditions, events and the execution of functions for positions from P ; $Et \subset \mathbb{N}$ is the set of integer energy inputs for performing actions and functions for transitions from T ; $F: (P \times X \times In \times Pb \times Ep \rightarrow) \cup (T \times Y \times In \times Pb \times Et \rightarrow P)$ - extended conditional incidence relation of transition positions; $S: (P \rightarrow X \times In \times Pb \times Ep) \cup (T \rightarrow Y \times In \times Pb \times Et)$ - extended correspondence of values of variable conditions, events, actions, functions, time intervals, probability coefficients to positions and transitions; $M_0: P \rightarrow \mathbb{N}$ - initial marking, $(M: P \rightarrow \mathbb{N}$ - function of current marking); $L: (T \times Y \times In \times Pb \times Et \rightarrow \{0, 1\})$ - transition predicate; $K: (((P \times X \times In \times Pb \times Ep) \rightarrow (P \times X \times In \times Pb \times Ep)) \cup ((T \times Y \times In \times Pb \times Et) \rightarrow (T \times Y \times In \times Pb \times Et)))$ is the function of modifying the values of variable conditions, events, actions, functions, time intervals, probability coefficients for positions and transitions.

Due to their asynchronous-event nature and token mechanism (parallel processes), multicomponent spatial decomposition of the PN can be performed for any subgraph of the PN graph, provided that the properties of the bipartite structure are preserved. Such a decomposition can formally be based on a two-component partition of the sets of positions and transitions, taking into account their incidence of the form:

$$S'(f)=(P', T', X', Y', In', Pb', Ep', Et', F', S', M_0', L', K'), \\ S''(f)=(P'', T'', X'', Y'', In'', Pb'', Ep'', Et'', F'', S'', M_0'', L'', K''), \quad (2)$$

where $P' \cup P'' = U$ и $P' \cap P'' = \emptyset$, $T' \cup T'' = U$ и $T' \cap T'' = \emptyset$, $X', X'' \subset X$, $Y', Y'' \subset Y$, $In', In'' \subset In$, $Pb', Pb'' \subset Pb$, $Ep', Ep'' \subset Ep$, $Et', Et'' \subset Et$, $M_0', M_0'' \subset M_0$, F' and F'' - narrowing the ratio F, S' и S'' - narrowing the correspondence S, L' and L'' - narrowing the predicate function L, K' and K'' - narrowing the function K .

As a result of decomposition, a spatial (S - Space) model is formed, defined by the set of all Petri subnets $S(f)_h^S \in S(f)^{S\wedge} = \cup_{h \in H} S(f)_h^S$ of the form $S'(f), S''(f)$ for the original PN $S(f)$, the structure of their connections in the composition, both the connections between the Petri subnets themselves, and the connections of the PSN with the external inputs and outputs of the composition — the inputs and outputs of the original PN $S(f)$. Additionally, functional-alphabetic relationships can be distinguished for inputs and outputs of PSN in accordance with the structure of their connections.

The resulting network decomposition model nS from $\forall S(f)_h^S \in S(f)^{S\wedge}$, which represents the spatial components of the DIS and their connections, has the form:

$$nS = (X, Y, S(f)^{S\wedge}, \alpha^\wedge), \quad (3)$$

where X is the input alphabet on the boundary nS ; Y is the output alphabet on the boundary nS ; $S(f)^\wedge$ is the set of component Petri subnets $\forall S(f)_h \in S(f)^\wedge$; α^\wedge is the set of functional-alphabetic correspondences (connections) between Petri subnets from $S(f)^\wedge$ to nS .

In nS , the operations of composition of functional PSN are used, which provide for parallel work, taking into account the markup function M . These are operations: serial connection $(S(f)_k \equiv S(f)_m)$, when the output positions $S(f)_k$ are the input positions for $S(f)_m$; parallel to the connection $S(f)_k \times S(f)_m$, when $S(f)_k$ and $S(f)_m$ have common input positions; feedback connections $(S(f)_k \equiv S(f)_m)$ when the output positions $S(f)_k$ are input positions for $S(f)_m$ and at the same time some output positions $S(f)_m$ are input positions for $S(f)_k$.

A multi-level hierarchical extended PN can be represented on the basis of a two-level hierarchical extended PN $2iS$ of the form:

$$2iS = (S(f), \cup_{i \in I} S(f)_i^p, \cup_{j \in J} S(f)_j^t, Sg_{is}), \quad (4)$$

where $S(f)$ is the highest PN from the upper level of the hierarchy; $S(f)^p = \cup_{i \in I} S(f)_i^p$ is the set of PSN of the lower hierarchy level that replace (with synchronization, translation) macro positions from $P' = \cup_{i \in I} p_i'$, where $P' \subset P$, for the PN $S(f)$ the upper level through the substitution of hierarchical correspondences χ^p, μ^p ; $S(f)^t = \cup_{j \in J} S(f)_j^t$ is the set of PSN s of the lower level of the hierarchy, replacing the macro positions from $T' = \cup_{j \in J} t_j'$, where $T' \subset T$, for the PN $S(f)$ of the upper level through substitution of hierarchical correspondences $\nu, Sg_{is} = \{\chi^{-p}, \chi^{p \rightarrow}, \nu^{-t}, \nu^{t \rightarrow}\}$ is the signature of the hierarchical correspondences themselves, in which χ^{-p} is the partial correspondence of the substitution of inputs to split macro positions from $P' = \cup_{i \in I} p_i'$ for the PN $t S(f)$ of the upper level at the entrances to new initial positions from the set $\cup_{i \in I} P_{S(f)_i^p}$ for PSN $S(f)^p = \cup_{i \in I} S(f)_i^p$ of the lower level; $\chi^{p \rightarrow}$ is the partial correspondence between the substitution of outputs from split macro positions from $P' = \cup_{i \in I} p_i'$ for the top-level PN $S(f)$ to the outputs from new end positions from the set $\cup_{i \in I} P_{S(f)_i^p}$ for PSN $S(f)^p = \cup_{i \in I} S(f)_i^p$ lower level; ν^{-t} is a partial correspondence between the substitution of inputs to split macro transitions from $T' = \cup_{j \in J} t_j'$ for the upper level PN $S(f)$ to the

inputs of new initial transitions from the set $\cup_{j \in T} T_{S(f)_j}$ for PSN $S(f)^i = \cup_{j \in T} S(f)_j^i$ of the lower level, v^{\rightarrow} is the partial correspondence of the substitution of the outputs from the split macro transitions from $T^i = \cup_{j \in T} t_j^i$ for the PN $S(f)$ of the upper level to the outputs of the new final transitions from the set $\cup_{j \in T} T_{S(f)_j}$ for PSN $S(f)^i = \cup_{j \in T} S(f)_j^i$ of the lower level.

As a result of decomposition, a temporary (τ - Temporal) model is formed, determined by the set of all PSN $\forall S(f)_h^T \in S(f)^T \wedge = \cup_{h \in H} S(f)_h^T$ included in it, in particular, of the form $S(f)_j^p \in S(f)^p$ or $S(f)_j^i \in S(f)^i$, by the structure of their correspondence-substitutions instead of the positions and transitions of PSN from $S(f)^T \wedge$, including the original PN $S(f)$, that is, the structure of auto-substitutions in the temporary composition. Additionally, functional-alphabetic relations can be distinguished for alphabets of conditions, events, actions, functions, in particular, inputs and outputs in accordance with their substitutions.

Obtained on the basis of the two-level model $2iS$ as a result of temporary hierarchical decomposition, the multi-level model iS from $\forall S(f)_h^T \in S(f)^T \wedge$, which represents the temporary components of the DIS and their relationships, has the form:

$$iS = (S(f)^T \wedge, Sg_{iS}), \quad (5)$$

where $S(f)^T \wedge = \cup_{h \in H} S(f)_h^T$ is the set of all PSN included in it, both with replaced positions / transitions and their substituting, $Sg_{iS} = \{\chi^{\rightarrow p}, \chi^{\rightarrow i}, v^{\rightarrow}, v^{\rightarrow i}\}$ is the signature of the hierarchical correspondences themselves, in which $\chi^{\rightarrow p}$ is the partial correspondence of the entries in the split macro positions from $P^i = \cup_{i \in P} p_i^i$ for some PSN $S(f)_i^T \in S(f)^T \wedge$ to the inputs of new initial positions from the set $\cup_{h \in H} P_{S(f)_h}$ for PSN $S(f)_h^{pT} \in S(f)^{pT} \wedge \subseteq S(f)^T \wedge$; $\chi^{\rightarrow i}$ - partial correspondence of substituting exits from split macro positions from $P^i = \cup_{i \in P} p_i^i$ for the PSN $S(f)_i^T \in S(f)^T \wedge$ to exits from new end positions from the set $\cup_{h \in H} P_{S(f)_h}$ for PSN $S(f)_h^{pT} \in S(f)^{pT} \wedge \subseteq S(f)^T \wedge$; v^{\rightarrow} is the partial correspondence between the substitution of inputs to split macro transitions from $T^i = \cup_{j \in T} t_j^i$ for some PSN $S(f)_j^T \in S(f)^T \wedge$ to the inputs of new initial transitions from the set $\cup_{h \in H} T_{S(f)_h}$ for PSN $S(f)_h^{iT} \in S(f)^{iT} \wedge \subseteq S(f)^T \wedge$; $v^{\rightarrow i}$ is the partial correspondence of substituting outputs from split macro transitions from $T^i = \cup_{j \in T} t_j^i$ for some PSN $S(f)_j^T \in S(f)^T \wedge$ to the exits from new finite transitions from the set $\cup_{h \in H} T_{S(f)_h}$ for PSN $S(f)_h^{iT} \in S(f)^{iT} \wedge \subseteq S(f)^T \wedge$. The hierarchy of PSN iS , formed by correspondences from Sg_{iS} on the set $S(f)^T \wedge$, can formally have a network structure and even include feedbacks, which should be justified by the object load of the model.

Let $S(f)^T \wedge = S(f)^S \wedge \cup S(f)^T \wedge$. As a result, the network hierarchical model obtained as a result of the combined spatial-temporal decomposition has the following form:

$$niS = (X, Y, S(f)^T \wedge, \alpha^T, Sg_{iS}), \quad (6)$$

The choice of a spatial-temporal decomposition of the initial PN model $S(f)$ for the DIS depending on its dimension, overall structural complexity, specific subject load and tuning suggests the existence of a multitude of solutions with a corresponding system of criteria that performs target structuring. The presented multidimensional spatial-temporal and hierarchical representations of PN models defined by the correspondences α^T and Sg_{iS} on the joint set $S(f)^T \wedge$ can give a general approach to the construction of such a set of solutions and decomposition criteria, to the choice of its variant.

From the formal point of view, the set of entities and relations of PN from $S(f)^T \wedge$, including their positions $P = \cup_{h \in H} P_h$, transitions $T = \cup_{h \in H} t_h$, chips $M = \cup_{h \in H} m_h$, the conditions $C = \cup_{h \in H} c_h$, events $E = \cup_{h \in H} e_h$ (as systems of conditions), actions $A = \cup_{h \in H} a_h$ (as systems of functions) and functions $F = \cup_{h \in H} f_h$, including inputs $X = \cup_{h \in H} x_h$ and outputs $Y = \cup_{h \in H} y_h$, where $X \subseteq C \cup E$ and $Y \subseteq A \cup F$, based on partitions or coverings, can be assigned, including multiple, to different spatial-temporal subsystems of the simulated DIS, that is, different submodels of PN from $S(f)^T \wedge$. Moreover, these coverings and partitions determine the nature of their interactions - strong connectedness, connectedness, weak connectedness, unconnectedness. In the case of only partitions, the spaces are not connected - not interacting or mutually independent.

In the general case, for relations $R = R_{c-e} \cup R_{e-a} \cup R_{a-f}$ between entities (Quintessence) from the sets $Q = C \cup E \cup A \cup F \cup X \cup Y$ of PSN $S(f)^T \wedge$ "event-conditions" R_{c-e} , "action-events" R_{e-a} , "action-functions" R_{a-f} is assigned to the $n-n$ type. It is possible to accept each submodel $S(f)_h \in S(f)^T \wedge$, whether it is the above network spatial PSN $S(f)_h^S \in S(f)^S \wedge$ or a temporary PSN substituting for a position or transition $S(f)_h^{pT} \in S(f)^{pT} \wedge$ or $S(f)_h^{iT} \in S(f)^{iT} \wedge$ of the corresponding DIS subsystem, as existing in a separate autonomous space-time or dimension. In this case, it is possible to speak of a multidimensional DIS model, and, in it, time decompositions $S(f)_h^{pT} \in S(f)^{pT} \wedge$ or $S(f)_h^{iT} \in S(f)^{iT} \wedge$ also form the corresponding hierarchies of temporal measurements similar to formal point of view of spatial dimensions.

In such a multidimensional model, on the one hand, the obtained formal, simple, and multiple submodel projections of the form $S(f)_h^S$ or $S(f)_h^{pT}$, $S(f)_h^{iT}$, decomposing the PN $S(f)$, in the form nS , iS , niS allow you to highlight simple and complex submodels of the corresponding subsystems of DIS. On the other hand, the inverse formal composition of submodels from $S(f)_h^S$ or $S(f)_h^{pT}$, $S(f)_h^{iT}$ of the DIS subsystems, presented as simple and complex projections, as a result forms the general compositional model nS , iS , niS for $S(f)$ with the degree of their interaction, determined by coverings and partitions of entities from Q and relations from R combined into composition of submodels.

Entities from Q and relations from R can be divided into sets with elements of the same multiplicity and form a hierarchy of static interactions of the submodels $Hi(S(f)^T \wedge)^{static}$, in which the ranks are determined by the multiplicity, as well as the correspondence relations α^T , Sg_{iS} . The dynamic interaction of submodels, determined by the statistics of activation of entities from Q and relations from R during the functioning of submodels from $S(f)^T \wedge$ and the model $S(f)$, can be obtained directly in the process of modeling the behavior of the DIS as a whole. The hierarchy $Hi(S(f)^T \wedge)$, additionally weighted by the dynamic statistics on the links, allows one to determine the structured quantitative measure / metric $Met(S(f)^T \wedge)$ for each submodel $S(f)_h \in S(f)^T \wedge$ of the form $Met(S(f)_h^T \wedge) = (Met(S(f)_h^T \wedge)^{static}, Met(S(f)_h^T \wedge)^{dynamic})$, consisting of static and dynamic components.

The metric $Met(S(f)^T \wedge)$ $Met(S(f)^T \wedge)$ admits ranking in the behavioral interaction of submodels from $S(f)^T \wedge$ $S(f)^T \wedge$ in the general complex model nS , iS , niS nS , iS , niS . The metric of behavioral interaction of the $Met(S(f)^T \wedge)$ $Met(S(f)^T \wedge)$ submodels, in turn, allows us to determine the so-called "near, middle, and far circles" of the interaction — often, infrequently, graph-space-time structures of its components (submodels from $S(f)^T \wedge$ $S(f)^T \wedge$) - Chains, Trees, Hammocks, Strongly Coupled Components SCC.

III. MODELS OF BEHAVIORAL TESTING OF DIS

The use of extended PN in the organization of behavioral online testing of DIS involves the recognition of characteristic fragments of the behavior of a reference PN in the working behavior of the checked PN, as a model of the project or the implementation of the DIS.

The class of checked properties Pr of the reference PN $S(f)$, for which the deviations of the tested SP $S(f)^\wedge$ are determined and the control model is determined, includes deviations of the incidence correspondences F^\wedge and S^\wedge from the reference correspondences F and S with the restriction $|P^\wedge| \leq |P|$ and $|T^\wedge| \leq |T|$. The error class of the PN $S(f)^\wedge$ is represented by the static part - its correspondences F^\wedge and S^\wedge , and the dynamic part - by its marking functions M^\wedge , predicates L^\wedge , modification of variables K^\wedge .

The model of behavioral testing cS has the form:

$$cS = (W^\wedge, Pr, Ci, Cp, Cf, Sg_{cS}), \quad (7)$$

where W^\wedge is the set of words of external (not structured by recognized positions and transitions) behaviour, which extends the incidence relation F , understood as the reachability relation on the unified set $P \cup T$; Pr is the checked properties based on the total incidence F ; $PrU = \{PrX \cup PrY\}$ is the checked properties based on the particular S included in F ; Ci - the identifying properties (identifiers of positions or transitions), for some $ci_{jk} \in Ci$ defined as two of the form $ci_{j_k p} = (p_{jk}, W_{jk})$, $W_{jk} = \cup_{j_{ki} \in I_k W_{j_{ki}}} \subset W_j$ identifiers of positions or $ci_{j_k t} = (t_{jk}, W_{jk})$, $W_{jk} = \cup_{j_{ki} \in I_k W_{j_{ki}}} \subset W_j$ identifiers of transitions for the reference $S(f)$ are uniquely incident to the corresponding positions of the p_{jk} and transitions of the t_{jk} , on the set of relations $\{\xi, \neg\xi, \varepsilon, \psi\}$ of compatibility, incompatibility, uncertainty and precedence (quasi order) are valid, taking into account the incidence of positions and transitions; Cp - the checked primitives, based on properties Pr and identifiers Ci ; Cf - recognized checked fragments of behaviour of reference PN $S(f)$, included the primitives Cp ; $Sg_{cS} = \{\alpha, \beta, \gamma\}$ - signature of operations: α - identification of positions or transitions; β - sameness of positions or transitions; γ - determinism of the behaviour of unmarked positions or transitions.

The network spatial model of check cnS for interacting in a system PSN has the form of a six:

$$cnS = (CS, node^\wedge, R_T^{-1}(node^\wedge), Tr_{T(node^\wedge)}, Sg_{cnS}, R_{(S(f)^\wedge)}, Tr_{(S(f)^\wedge)}), \quad (8)$$

where: $CS = \cup_{h \in HC} S_h$ - a lot of control models for network spatial memory bandwidths of the previously given form; $node^\wedge$ - the set of selected nodes (selected pairs of adjacent positions and transitions for the PSN $S(f)^\wedge$, in which there is a convergence-divergence of chip flows and selected entities from Q and R represent the behavior of implementation (control) and transportation (observation) inside and at the boundaries of the composition nS ; $R_T^{-1}(node^\wedge)$ is the set of submodels of the implementation of behavior (in the form of minimized PSN) on each node $\forall node_h \in node^\wedge$ of the composition nS - the system PN $S(f)$ from its common inputs through the reverse (to common inputs) simple (without repetitions) graph network spatial structures from the PSN of the form $S(f)^\wedge \hat{R}^{-1}(node_h) \subseteq S(f)^\wedge$ with the nodes $node_{RT^{-1}(node_h)} \subseteq node^\wedge$ selected in them; $Tr_{T(node^\wedge)}$ is the set of submodels of behavior transportation (in the form minimized CSP) from each of the nodes $\forall node_h \in node^\wedge$ of the composition nS -

system PSN $S(f)$ to its common outputs through direct (common outputs) simple graph network spatial structures from the CSP of the form $S(f)^\wedge \hat{R}^{-1}(node_h) \subseteq S(f)^\wedge$ with nodes selected in them $node_{RT^{-1}(node_h)} \subseteq node^\wedge$; $Sg_{cnS} = \{\circ^Y, \circ^X, \times^Y, \times^X, *^Y, *^X\}$ - the signature of the inter-component, between the PSN from $S(f)^\wedge$ through the corresponding nodes from $node^\wedge$ network direct and reverse operations compositions - serial \circ^Y, \circ^X , parallel \times^Y, \times^X , with feedback $*^Y, *^X$ - for the behavior represented by constraints (for control and observation) of the PSN; $R_{(S(f)^\wedge)}$ - the set of submodels of the implementation of behavior (in the form of minimized PSN) at the inputs (input nodes $node_{S(f)^\wedge}^{in} \subseteq node^\wedge$ of each of the PSN $\forall S(f)_h \in S(f)^\wedge$ implemented through the nodes $\forall R_T^{-1}(S(f)_h) \subseteq R_T^{-1}(node^\wedge)$; corresponding to $S(f)_h$; $Tr_{(S(f)^\wedge)}$ - sets of submodels of behavior transportation (in the form of minimized PSN) from the outputs (output nodes $node_{S(f)^\wedge}^{out} \subseteq node^\wedge$ of each PSN $\forall S(f)_h \in S(f)^\wedge$ transported through the nodes $\forall Tr_{T(S(f)_h)} \subseteq Tr_{T(node^\wedge)}$ corresponding to $S(f)_h$.

The network check model cnS accepts, as input, component check models CS and limits them to the conditions of realizability (controllability) and transportability (observability). In the model cnS , in addition to the set of component CS check models for all component PSN $\forall S(f)_h \in S(f)^\wedge$, the construction of all proper implementation models from $\forall R_{(S(f)_h)} \in R_{(S(f)^\wedge)}$ and transportation $\forall Tr_{(S(f)_h)} \in Tr_{(S(f)^\wedge)}$ in the network composition nS and reuse of node behavior implemented by $R_T^{-1}(node^\wedge)$ and transported by $Tr_{T(node^\wedge)}$ in node $node^\wedge$ composition nS is emphasized.

The set of output words determined on the basis of $Tr_{T(node^\wedge)}$, $Tr_{(S(f)^\wedge)}$ and transported by the network nS , respectively, from the outputs of nodes from $node^\wedge$ and PSN from $S(f)^\wedge$, are based on a number of additional models, in particular, models of specially generated check graphs $\forall G_{S(f)_h} \in G_{S(f)}$ for all $\forall S(f)_h \in S(f)^\wedge$. When solving online testing problems, the $2iS$ hierarchy imposes conditions for inheriting check behavior in hierarchical transitions from $2iS$ under replacement mappings for detailed PSN from $\cup_{i \in I} S(f)_i^p$, $\cup_{j \in J} S(f)_j^t$ instead of the corresponding macropositions and macrotransitions of the system PN $S(f)$.

Thus, the organization of the 2-hierarchy of check primitives and, as a result, the behavioral working control performed when they are covered is possible provided that the verifiable properties and position identifiers in the replacement PSN from the $\cup_{i \in I} S(f)_i^p$, $\cup_{j \in J} S(f)_j^t$ in the signature $Sg_{iS} = \{\chi^{\rightarrow p}, \chi^{\rightarrow t}, v^{\rightarrow p}, v^{\rightarrow t}\}$ hierarchical mappings $\chi^{\rightarrow p}, \chi^{\rightarrow t}$ of the positions and mappings $v^{\rightarrow p}, v^{\rightarrow t}$ transitions from $\cup_{i \in I} S(f)_i^p$, $\cup_{j \in J} S(f)_j^t$ of each two-level $2iS$ hierarchy. This condition restricts the sets $\cup_{i \in I} S(f)_i^p$, $\cup_{j \in J} S(f)_j^t$ of junior implementers of the memory bandwidth in the hierarchical maps from Sg_{iS} to be valid for storing checked properties and identifiers. In this case, the set of preserving hierarchical transitions forms five compatible hierarchies derived from the initial hierarchy $2iS = (S(f), \cup_{i \in I} S(f)_i^p, \cup_{j \in J} S(f)_j^t, Sg_{iS})$, - the reference checked properties $2iPr$, position identifiers $2iCi$, control primitives $2iCp$, fragments of fixed recovered behavior $2iCf$, fragments of fixed unrecovered behavior $2iW^\wedge$:

$$\begin{aligned} 2iPr &= (Pr, \cup_{i \in I} Pr_i^p, \cup_{j \in J} Pr_j^t, Sg_{iPr}), \\ 2iCi &= (Ci, \cup_{i \in I} Ci_i^p, \cup_{j \in J} Ci_j^t, Sg_{iCi}), \\ 2iCp &= (Cp, \cup_{i \in I} Cp_i^p, \cup_{j \in J} Cp_j^t, Sg_{iCp}), \\ 2iCf &= (Cf, \cup_{i \in I} Cf_i^p, \cup_{j \in J} Cf_j^t, Sg_{iCf}), \\ 2iW^\wedge &= (W^\wedge, \cup_{i \in I} W_i^p, \cup_{j \in J} W_j^t, Sg_{iW^\wedge}), \end{aligned} \quad (9)$$

where $Sg_{iP}=Sg_{iS(P)}\subseteq Sg_{iS}$, $Sg_{iC}=Sg_{iS(C)}\subseteq Sg_{iS}$, $Sg_{iCp}=Sg_{iS(Cp)}\subseteq Sg_{iS}$, $Sg_{iCf}=Sg_{iS(Cf)}\subseteq Sg_{iS}$, $Sg_{iW}=Sg_{iS(W)}\subseteq Sg_{iS}$ is the set of subsets Sg_{iS} , that is, the map $Sg_{iCS}=\{Sg_{iS}, Sg_{iP}, Sg_{iC}, Sg_{iCp}, Sg_{iCf}, Sg_{iW}\}$ is a special covering of the Sg_{iS} map, including the Sg_{iS} map itself.

Then for defined two-level check hierarchy model of two-level check $2icS$ has the form:

$$2icS = (cS, \cup_{i \in I} cS_i^P, \cup_{j \in J} cS_j^T, Sg_{iCS}). \quad (10)$$

The set of retaining hierarchical transitions forms the top five compatible hierarchies derived from the initial hierarchy $iS=(S(f)^T, Sg_{iS})$, iPr is the hierarchy for the reference checked properties $Pr^T = \cup_{h \in H} Pr_h^T$, iCi is the hierarchy for the identifiers of the positions $Ci^T = \cup_{h \in H} Ci_h^T$, iCp - hierarchy for control primitives $Cp^T = \cup_{h \in H} Cp_h^T$, iCf - hierarchy for fragments of fixed restored behavior $Cf^T = \cup_{h \in H} Cf_h^T$, iW - hierarchy for fragments of fixed unrestored behavior $W^T = \cup_{h \in H} W_h^T$.

$$\begin{aligned} iPr &= (Pr^T, Sg_{iP}), & iCi &= (Ci^T, Sg_{iC}), \\ iCp &= (Cp^T, Sg_{iCp}), & iCf &= (Cf^T, Sg_{iCf}), \\ iW &= (W^T, Sg_{iW}). \end{aligned} \quad (11)$$

Obtained on the basis of the $2icS$ two-level behavioral control model as a result of temporary hierarchical decomposition, the ciS multi-level behavioral control model for the iS hierarchy from $\forall S(f)_h^T \in S(f)^T$, which represents the corresponding time components $cS^T = \cup_{h \in H} cS_h$ of the multilevel RIS behavioral control and their hierarchical relationships in the Sg_{iCS} mapping signature is:

$$ciS = (cS^T, Sg_{ciS}). \quad (12)$$

where $cS^T = \cup_{h \in H} cS_h$ is the set of all hierarchical submodels of behavioral control included in it for $S(f)^T = \cup_{h \in H} S(f)_h^T$, firstly, for submodels with replaceable positions/transitions and previous detailed properties, by identifiers, primitives, fragments, and secondly, for submodels that replace positions / transitions and contain new detailed properties, identifiers, primitives, fragments. The hierarchy of multilevel behavioral check ciS , as well as the hierarchy of PN iS , formed by matches from Sg_{iCS} on the set cS^T ,

can formally have a network structure and include feedbacks, which should be justified by the subject load of the control model.

Thus, for the iS hierarchy, a hierarchical model of multilevel behavioral testing ciS is formed.

Based on the hierarchical network model niS , a hierarchical network model of behavioral check $cnis$ is formed as follows:

$$cnis = (cS^S, cS^T, node^A, R_T^{-1}(node^A), Tr_{T(node^A)}, Sg_{cns}, R_{(S(f)^A)}, Tr_{(S(f)^A)}, Sg_{cis}). \quad (13)$$

In the online testing, analysis of the hierarchical mappings $\{\chi^p, \chi^q, v^x, v^y\}$ of the highest PN $S(f)$ and any component PSN from $S(f)^+ = S(f) \setminus ((\cup_{i \in P_i} \cup_{j \in T_j}) \cup (\cup_{i \in S(f)_i^P} \cup_{j \in S(f)_j^T}))$ $2iS$ hierarchy is performed for relations quasi-order $\psi^+ = \psi \setminus (\cup_{i \in I} \psi_i) \setminus (\cup_{j \in J} \psi_j)$, compatibility $\xi^+ = \xi \setminus (\cup_{i \in I} \xi_i) \setminus (\cup_{j \in J} \xi_j)$, representing the higher and lower levels of synchronization behavior of the hierarchical transition for $S(f)^+$.

IV. EVALUATION OF BEHAVIORAL TESTING MODELS OF DIS

Representation of the Petri network $S(f)$, where $|P|=n_p$, $|T|=n_t$, $n=n_p+n_t$, $|X|=m$, $|Y|=L$, in the memory of the monitoring system using list structures requires for the upper limit of the total number of conditional fields:

$$c_{iS}^{max} = (n_i(4n_p+3L+4)+n_p(3m+2)) + (\sum_{i \in I} n_{i_i}(4n_{p_i}+3L_i+4)+n_{p_i}(3m_i+2)) + (\sum_{j \in J} n_{j_j}(4n_{p_j}+3L_j+4)+n_{p_j}(3m_j+2)). \quad (14)$$

The complexity of the check analysis of the PN $S(f)$ is determined by the upper bound:

$$\begin{aligned} c_{ciS}^{max} &= n_i(4n_p+3L+4)+n_p(3m+2)+2n_p n_i(n_i-1)+2(2Lm n_p n_i)^{n_i}- \\ &- 3)+(n_i-1)(n_p n_i)! + (\sum_{i \in I} n_{i_i}(4n_{p_i}+3L_i+4)+n_{p_i}(3m_i+2)+ \\ &+ 2)+2n_{i_i} n_{p_i}(n_{i_i}-1)+2(2L_i m_i n_{p_i} n_{i_i})^{n_{i_i}-3}+(n_{i_i}-1)(n_{p_i} n_{i_i})! + \\ &+ (\sum_{j \in J} n_{j_j}(4n_{p_j}+3L_j+4)+n_{p_j}(3m_j+2)+2n_{j_j} n_{p_j}(n_{j_j}-1)+ \\ &+ 2(2L_j m_j n_{p_j} n_{j_j})^{n_{j_j}-3}+(n_{j_j}-1)(n_{p_j} n_{j_j})!). \end{aligned} \quad (15)$$

Upper computational complexity and length of the online testing for the Determination (solid) and Evolutional (dashed) methods in cases of simple Petri net (●) and hierarchical Petri net (▲) are presents in Fig. 1.

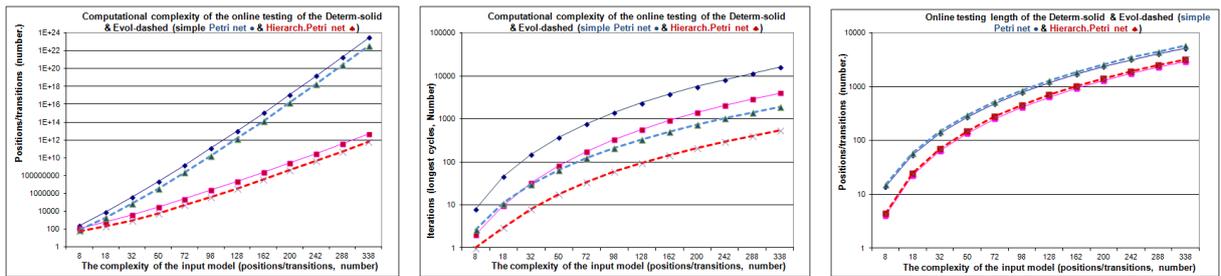


Figure 1 - Computational complexity and length of testing of the Determ-solid & Evol-dashed (simple PN ● & Hierarch.PN ▲)

Experimental check of the procedures and programs of behavioral testing were carried out for component and decomposition of DIS, the results are presented in Tab. 1.

Table 1 – Experimental values of computational complexity of check

Object	Degree of decomposition	Input complexity	Complexity of check
Module BSAC	8	362	64980
Module TVS	2	50	4323
Module IP/IPSec	3	115	71871
Module of Naming	2	146	38150450
Module of Encaps.	3	180	1572123

The comparison of the work of online testing programs based on automata-deterministic and Petri-evolutionary

methods for dynamic DIS of onboard automated control systems (BSAC) and terminal video surveillance (TVS) confirmed a) decrease in computational complexity of online testing; b) reducing timer time of the check with preservation of c) the length of the check and their d) completeness.

V. CONCLUSIONS

The paper presents the results of the development of the method of behavioral online testing of distributed information systems based on a special model of behavioral control of extended Petri nets and characterized by the features of the “wave” evolutionary parallelism of the “background” control analysis. A special model of

behavioral control of extended Petri nets is based on determining the compliance of the reference and verifiable extended Petri nets, representing respectively the reference and verifiable components of DIS. The use of the model made it possible to determine the basic conditions for constructing a check method applied both at the system level and at the component level.

Decomposition increases the flexibility of the organization of behavioral workers control by taking into account the features of DIS. The greatest reduction was achieved on the components of the DIS of special behavior, in particular, with partial definiteness of model functions.

REFERENCES

- [1] G. Coulouris, J. Dollimore, T. Kindberg, G. Blair, "Distributed Systems: Concepts and Design," 5th ed., Boston: Addison-Wesley, 2011.
- [2] M. Van Steen, A. S. Tanenbaum Maarten, "A brief introduction to distributed systems," *Computing*, 2016, no. 98 (10), pp. 967-1009.
- [3] H. Benítez-Pérez, J. Ortega-Arjona, P. Méndez-Monroy, E. Rubio-Acosta, O. Esquivel-Flores, "Distributed Systems Modelling: In: Control Strategies and Co-Design of Networked Control Systems, Modeling and Optimization in Science and Technologies, vol 13, Springer, Cham, pp. 41-82, 2019.
- [4] B. Großwindhager, A. Rupp, M. Tappler, M. Tranninger, S. Weiser, B. K. Aichernig, C. A. Boano, M. Horn, G. Kubin, S. Mangard, M. Steinberger, K. Römer, "Dependable Internet of Things for Networked Cars," *International Journal of Computing*, 2017, vol. 16, Issue 4, pp. 226-237.
- [5] R. E. Hiromoto, "Parallelism and Complexity of a Small-World Network Model," *International Journal of Computing*, 2016, vol. 15, Issue 2, pp. 72-83.
- [6] Y. Abdeddaïm, M. Dorin, "Probabilistic Schedulability Analysis for Fixed Priority Mixed Criticality Real-Time Systems," *Design, Automation and Test in Europe - DATE 2017, Lausanne, Switzerland*, pp. 596-601, 2017.
- [7] B. Alahmad, S. Gopalakrishnan, "Isochronous execution models for mixed-criticality systems on parallel processors," *WiP, RTSS*, 2017, pp. 354-356.
- [8] J. Stuart, P. Norvig, "Artificial Intelligence: a Modern Approach, Prentice-Hall," A Simon & Schuster Company Englewood Cliffs, New Jersey, 2010.
- [9] R. Poli, W. B. Langdon, N. F. McPhee, J. R. Koza, "A Field Guide to Genetic Programming," *Creative Commons Attribution-NonCommercial No Derivative Works 2.0 UK: England*, March 2008.
- [10] S. Manaffam, "Stability and Control in Complex Networks of Dynamical Systems," *Dissertation*, 2015. <https://stars.library.ucf.edu/etd/692>.
- [11] W. Schamai, "Model-Based Verification of Dynamic System Behavior against Requirements Method, Language, and Tool," *Linköping University SE-581 83 Linköping, Sweden Linköping*, 2013.
- [12] Z. Zhou, H. Wang, P. Lou, "Manufacturing intelligence for industrial engineering: methods for system self-organization, learning, and adaptation," *Engineering Science Reference*, 2010.
- [13] D. A. Kanhere, S. J. Raja Multi-Agent Systems," *A survey. IEEE Access*. 1-1. 10.1109/ACCESS.2018.2831228, 2018.
- [14] R. Cohen, M. Schaekermann, S. Liu, M. Cormier, "Trusted AI and the Contribution of Trust Modeling in Multiagent Systems," *AAMAS, Montréal, Canada*, pp. 1644-1648 2019.
- [15] Y. Shoham, K. Leyton-Brown, "Multiagent systems. Algorithmic, Game-Theoretic, and Logical Foundations, Cambridge University Press, 2009.
- [16] L. Callebert, D. Lourdeaux, J.-P. Barthès, "A Trust-based Decisionmaking Approach Applied to Agents in Collaborative Environments," *International Conference on Agents and Artificial Intelligence, Science and Technology Publications*, pp. 287-295, 2016.
- [17] V. Kharchenko, A. Gorbenko, V. Sklyar, C. Phillips, "Green Computing and Communications in Critical Application Domains: Challenges and Solutions," *IX International Conference of Digital Technologies, Zhilina, Slovak Republic*, 2013, pp. 191-197.
- [18] V. Hahanov, E. Litvinova, V. Obrizan, W. Gharibi, "Embedded method of SoC diagnosis, *Elektronika in Elektrotechn*, no 8, pp. 3-8, 2008.
- [19] A. Drozd, M. Drozd, V. Antonyuk, "Features of Hidden Fault Detection in Pipeline Components of Safety-Related System," *CEUR Workshop Proceedings*, 2015, vol. 1356, pp. 476-485.
- [20] A. Jason, "The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice," *Syngress*, 2014.
- [21] A.V. Drozd, M.V. Lobachev, "Efficient On-line Testing Method for Floating-Point Adder," *Proc. Design, Automation and Test in Europe. Conference and Exhibition 2001 (DATE 2001), Munich, Germany, 2001*, pp. 307-311. DOI: 10.1109/DATE.2001.915042
- [22] A. Drozd, M. Drozd, O. Martynyuk, M. Kuznietsov, "Improving of a Circuit Checkability and Trustworthiness of Data Processing Results in LUT-based FPGA Components of Safety-Related Systems," *CEUR Workshop Proceedings*, 2017, vol. 1844, pp. 654-661.
- [23] K. Zashcholkin, O. Ivanova, "The control technology of integrity and legitimacy of LUT-oriented information object usage by selfrecovering digital watermark," *CEUR Workshop Proceedings*, 2015, vol. 1356, pp. 498-506.
- [24] K. Zashcholkin, O. Ivanova, "LUT-object integrity monitoring methods based on low impact embedding of digital watermark," *14th International Conference "Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering*, 2018, pp. 519-523.
- [25] C. H. Hong, I. Spence, D. S. Nikolopoulos, "Gpu virtualization and scheduling methods," *ACM Comput. Surv.*, 2017, vol 50(3), pp. 1-37.
- [26] M. Komar, A. Sachenko, V. Kochan, V. Ababii, "Improving the Security of Intrusion Detection System," *International Conference on Development and Application Systems, Suceava, Romania*, 2016, pp. 19-21.
- [27] P. Srivastava, Km Baby, "Automated Software Testing Using Metaheuristic Technique Based on an Ant Colony Optimization," *International Symp. on Electronic System Design, Bhubaneswar*, 2010, pp. 235-240.
- [28] J. Wegener, A. Baresel, H. Sthamer, "Evolutionary test environment for automatic structural testing," *Information and Software Technology*, 2001, no 43, pp. 841-854.
- [29] M. S. Phadoongsidhi, K. K. Saluja, "Logic Crosstalk Delay Fault Simulation in Sequential Circuits," *International Conference on VLSI Design, Kolkata, India*, 2005, pp. 820-823.
- [30] Yu. A. Skobtsov, V. Yu. Skobtsov, "Evolutionary test generation methods for digital devices," *Design of Digital Systems and Devices [eds.: M.Adamski et al.]*, Berlin: SpringerVerlag, 2011, pp. 331-361.
- [31] A. Sugak, O. Martynyuk, A. Drozd, "Models of the Mutation and Immunity in Test Behavioral Evolution, 8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Warsaw, Poland, 2015, pp.790-795.
- [32] O. Martynyuk, A. Sugak, D. Martynyuk, O. Drozd, "Evolutionary Network Model of Testing of the Distributed Information Systems," *9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Bucharest, Romania*, 2017, pp. 888-893.
- [33] M. Grindal, "Handling Combinatorial Explosion in Software Testing," *Printed by LiU-Tryck, Linköping*, 2007.
- [34] IEC 61508-1:2010. *Functional Safety of Electrical / Electronic / Programmable Electronic Safety Related Systems – Part 1: General requirements*. Geneva: International Electrotechnical Commission, 2010.
- [35] V. B. Kudryavtsev, I. S. Grunskii, V. A. Kozlovskii, "Analysis and synthesis of abstract automata," *Journal of Mathematical Sciences*, September 2010, vol. 169, Issue 4, pp. 481-532.
- [36] D. B. Achim, W. Burkhart, "On Theorem Prover-based Testing," *Formal Aspects of Computing*, 2012, no. 25 (5), pp. 683-721.
- [37] A. Drozd, J. Drozd, S. Antoshchuk, V. Antonyuk, K. Zashcholkin, M. Drozd, O. Titomir, "Green Experiments with FPGA," in book: *Green IT Engineering: Components, Networks and Systems Implementation, V. Kharchenko, Y. Kondratenko, J. Kacprzyk, Eds., Vol. 105*. Berlin: Springer, 2017, pp. 219-239. DOI: 10.1007/978-3-319-55595-9_11
- [38] L. Gomes, J. Fernandes, "Behavioral Modeling for Embrdded Systems and Technologies: Applications for Design and Implementation,," *InformatIon Science Reference, Hershey, New York*, 2010.
- [39] I. Atamanyuk, Y. Kondratenko, V. Shebanin, "Calculation Methods of the Prognostication of the Computer Systems State under Different Level of Information Uncertainty," *CEUR Workshop Proceedings*, 2016, vol. 1614, pp. 292-307.
- [40] J. Drozd, A. Drozd, S. Antoshchuk, A. Kushnerov, V. Nikul, "Effectiveness of Matrix and Pipeline FPGA-Based Arithmetic Components of Safety-Related Systems," *8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Warsaw, Poland*, 2015, pp. 785-789.
- [41] A. Drozd, M. Drozd and M. Kuznietsov, "Use of Natural LUT Redundancy to Improve Trustworthiness of FPGA Design," *CEUR Workshop Proceedings*, 2016, vol. 1614, pp. 322-331.

Development Of Method For Automation Of SPICE Models Generation

Melikyan Vazgen Sh., Martirosyan Meruzhan K.
Synopsys Armenia Educational Department
Synopsys Armenia CJSC
Yerevan, Armenia
vazgenm@synopsys.com, meruzha@synopsys.com

Abstract— A new method has been developed to obtain precise spice models of transistors that can be used in any type of research and training projects. The transistor's input and output characteristics have been received that have been compared with the characteristics of the foundry models. Also, Monte Carlo SPICE models were developed for all types of transistors, that are used during AOCV and POCV model's development processes.

Keywords— spice model, nanoscale transistor, aocv, pocv, montecarlo model, objective function

I. INTRODUCTION

Different types of discrete elements (eg, capacitance, resistance, transistor, etc.) are used when designing a board and interconnections are made between them. As a result, a circuit is made of different elements, after which this circuit is checked. If the circuit operates in line with the pre-set requirements, the work is over; otherwise individual components are replaced until the circuit behaves as designed and actually meets the specifications. However, such work cannot be done in projects based on integrated circuits (IC), the reason being the small size of the elements and the simultaneous production. In other words, any postproduction changes in the IC are impossible to make. The solution of the problem is to create virtual environment and models operating in that environment. Examples of such models are SPICE models that are used by the largest companies specializing in IC design. Various projects based on these models can be implemented – from the smallest-scale (e.g. an operational amplifier) to the largest-scale ones (e.g. i7 processor). However, these models are confidential and accessible to a very small number of researchers, who actually work in such companies. Thus, there arises the necessity to create models which will be as close as possible to real models and will be available to anyone who wishes to do research using quite accurate models of transistors. The basis of this work is the development of the method that will make creating such models possible.

II. NANOSCALE TRANSISTORS

The development of ICs has always been accompanied by the scaling of their elements. However, the scaling has been performed in a disproportionate manner for various reasons, that is, the size of the different parts of the IC has been reduced not with the same coefficients. This has resulted in deterioration of the parameters of the primary elements of the ICs, namely, the gates and hence transistors. In particular, in the case of 28 nm and 22 nm technologies, the basic parameters of the flat MOS transistors used in the IC have inadmissibly deteriorated, and further scaling was only possible using a new type of transistors. That is why new types of transistors, FinFET [1] (Figure 1), began to be used in the ICs made using the 14 nm and smaller-scale technological

processes. The principal difference of the latter from the flat transistors is in that it has a three-dimensional "fin", around

which a three-dimensional channel is formed. Due to this fact, the main parameters of the transistor, including leakage current and inertial properties, are greatly improved.

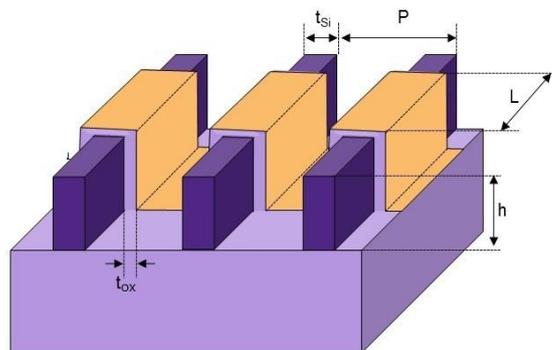


Fig. 1. The view of FinFET transistor

It is due to the above-mentioned properties that the further development of the nanoscale ICs is closely linked to FinFETs. According to various estimates, the IC market built with FinFET will have an annual growth of around 42% up to 2021. That is why the method developed during this work has been used to obtain FinFET-type transistor models.

III. MONTE CARLO MODEL

The Monte Carlo SPICE model is used that relies on repeated random sampling and statistical analysis to compute the results. This method of simulation is very closely related to random experiments, experiments for which the specific result is not known in advance. In this context, Monte Carlo simulation can be considered as a methodical way of doing so-called what-if analysis [2]. Based on that model On-Chip Variation (OCV) tables are developed.

In Fig. 2 is shown the simple Monte Carlo Mathematical model, where the model depends on number of input parameters and these parameters in turn depend on various external factors. When input parameters process through the mathematical formulas in the model, they result in one or more outputs.



Fig.2. Monte Carlo mathematical model

During manufacturing process many variables are used. Some of these variables are consistent for the whole process, some of them are consistent across a single wafer and some vary from wafer to wafer but are consistent across a

chip. Also, there are variables often observed in a single chip. This so called on-chip-variation (OCV) may come from mask alignment, etching process, and optical proximity correction. Therefore, two instances of the same cell on the same chip may have different timing characteristics.

Every cell has minimum and maximum delay due to OCV tables. According to this table static timing analyzer for setup time check applies minimum delay to clock path and maximum delay for data path. For hold time check it applies maximum delay to clock path and minimum delay for data path. The OCV's level of precision is low. To improve the accuracy of design timing analysis, one must apply Advanced On-Chip Variation (AOCV) or Parametric On-Chip Variation (POCV) tables.

A. Advanced On-Chip Variation

AOCV analysis reduces unnecessary pessimism by taking the design methodology and fabrication process variation into account. AOCV [3] determines derating factors based on cell type, metrics of path logic depth and the physical distance traversed by a particular path. The number of transistors and connection between them in each logic gate are different, therefore the variations for each cell are different. A longer path that has more gates tends to have less total variation because the random variations from gate to gate tend to cancel each other out. A path that spans a larger physical distance across the chip tends to have larger systematic variations. AOCV is less pessimistic than a traditional OCV analysis, which relies on constant derating factors that do not take path-specific metrics into account.

While performing register to register timing analysis, AOCV methodology finds the bounding box (Fig. 3) containing the sequential registers, clock buffers between two sequential registers and all the data cells. Now within a unit distance, if the path depth increases, the AOCV derate decreases due to cancelling of random variations. However, if the distance increases, AOCV derate increases due to increase in the systematic variations. These variations are modeled in form of a Look Up Table (LUT).

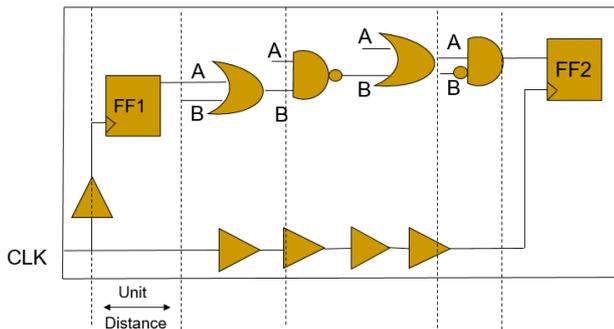


Fig.3. Bounding box creation for AOCV

B. Parametric On-Chip Variation

POCV models the delay of an instance as a function of a variable that is specific to the instance. That is, the instance delay is parameterized as a function of the unique delay variable for the instance. POCV [3] provides the following:

- Statistical single-parameter derating for random variations
- Single input format and characterization source for both AOCV and POCV table data
- Non-statistical timing reports

- Limited statistical reporting (mean, sigma) for timing paths

IV. DEFINITION OF PROBLEM OF AUTOMATED DEVELOPMENT OF SPICE MODEL OF TRANSISTOR

Two types of problems have to be solved. Since the transistor model contains several hundreds of parameters, we can conclude that some of the parameters have a greater impact on the behavior of the model than the others. Therefore, the first task is to classify the parameter settings by priority. After the parameters are classified, it is necessary to define the objective function, i.e. to define a function that must be found during solving the problem. And since the model parameters cannot take on arbitrary values, restrictions should also be noted, which must be met to find the objective function.

A. Parameters classification by priority

First of all, it is necessary to compile a list of parameters that will be included in the model. Two approaches can be used in obtaining the list of parameters: first, taking all the transistor parameters that the spice model can include. Secondly, use similar models created by Arizona State University (ASU) [4] or some of the models created for a different technology by Synopsys Armenia Educational Department (SAED) [5]. During this work, the latter approach was used in choosing the list of model parameters.

Let us call the selected parameters a_i where i is the natural number in the range $[1, N]$ and N is equal to the number of parameters included in the model. Let us take the following function:

$$y_j = f_j(a_i) \quad (1)$$

where:

$$j = \{P, T, I\},$$

P - dynamic and static power,

T - time parameters (rising transition, falling transition and propagation delay),

I - the current flow through the transistor.

For all values of j , one needed to count all Δy_j functions,

where:

$$\Delta y_j = \Delta f_j(\Delta a_i) = f_j(a_{i2} - a_{i1}) \quad (2)$$

$$a_{i2} = a_{i1} + \Delta a_i, \quad (3)$$

Δa_i - minimum permissible values of parameter deviation, which comes from the model.

Let us set another parameter:

$$K(a_i) = \prod \Delta f_j(a_i) \quad (4)$$

$K(a_i)$ - deviation value for each parameter.

So, the a_i parameter for which the value of (4) will be the largest is going to be the parameter that has the greatest impact on the behavior of the model.

Accordingly, the classification of the parameters by priority can be determined by the following algorithm:

Algorithm 1.

1. Calculate all (2)s, where $j=\{P,T,I\}$, and i the natural number from in the range[1, N].
2. Calculate all (4)s.
3. Summarize the results of (4)s in the associative array (such an array that has a key and a value, where the key indicates the model parameter name, and the value indicates the parameter value).
4. Make sorting according to array values, from big to small.
5. Take the keys of the sorted array.

The keys of the received array are the names of transistor parameters, sorted by impact. Thus, the first parameter affects the model to the largest extent and the last parameter affects to the smallest extent.

B. Objective Function

The goal is to find the minimum values of (2) for each a_i .

$$\Delta F_i = | F_{i2} - F_{i1} | \tag{5}$$

where:

$$i = \{P, T, I\},$$

F_{i2} - describes the behavior of the foundry model,

F_{i1} - describes the behavior of the model acquired during this work.

Thus, it is necessary to get the values of a_i parameters where the value of the F_{i1} function would be as close as possible to the value of the F_{i2} function.

C. Constraints

The function (5) takes on minimum value in the case of a certain value of a_i parameter, but at the same time this parameter cannot take on any value because the model does not allow it. According to the last statement:

$$a_{imin} \leq a_i \leq a_{imax} \tag{6}$$

$$a_i = a_{imin} + n * \Delta a_i \tag{7}$$

where:

a_{imin} - the minimum value of a parameter,

a_{imax} - the maximum value of a parameter,

Δa_i - sets the minimum step of or change in the parameter,

n – natural number.

As a result, the problem can be solved using the following algorithm:

Algorithm 2.

1. Select a nominal value of the a_i parameter (for example, this value can be taken from the model created by ASU).
2. Modify the value of the received parameter to the top and bottom, by following the restrictions until the minimum value of (5) is reached in the given parameter.
3. Repeat the same for the other parameters.
4. Do the same for as long, or for as many parameters, until the following takes place:

$$\Delta F_i / F_{i2} * 100\% \leq B \tag{8}$$

where:

B - permissible deviation.

5. According to the last point all the parameters, that satisfy the condition are selected to be the parameters of the new model. The parameters that do not satisfy the condition are not selected, but the ones which are obligatory for models to have are assigned default values and used in model development.

Thus, filling the values of a_i parameters in the transistor model using the Algorithm 2, we will get a model that is very close to the produced (foundry) model with its parameters.

V. APPLICATION OF ALGORITHM IN THE PROCESS OF OBTAINING OF THREE-DIMENSIONAL TRANSISTOR MODELS

Curves describing transistor parameters were constructed. Then, Algorithm 1 was used for classifying the parameters by priority. Different values of P, T, and I were taken from scientific papers [6-8]. Applying the interpolation algorithms to those values enabled transistor-describing curves to be obtained (input characteristics, output characteristics, time dependence on the transistor size, etc.). Then a_i nominal values were taken from the model created by the ASU, after which Algorithm 2 was used for finding the optimal dimensions of the transistor parameters. This process lasted until the B value equaled to 5%, which is considered to be permissible deviation. Fig. 4 shows how the curves of input characteristics of the model obtained with the algorithms approach to the curves of input characteristics of the foundry model.

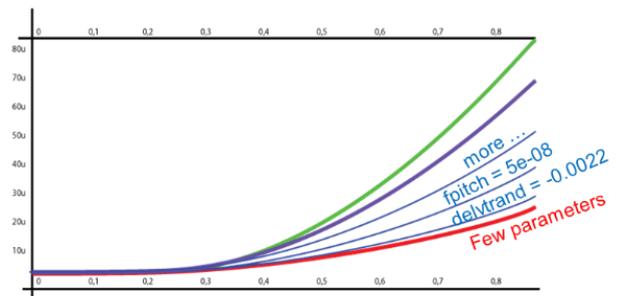


Fig. 4. Input characteristic of the model is approaching to foundry model's input characteristic

In the figure the green curve describes the foundry model, and the other colors describe the processed model (the model

under development), which is approaching the foundry model due to changes in the parameter values. The same thing has happened to the rest of the parameters.

It is common knowledge that transistors are exposed to certain technological deviations in the production process. These deviations cause transistors to operate faster or slower. An example of technological discrepancy is the acquisition of different lengths of the transistor's channel length. If a transistor's channel length will be greater than it was designed, it will work slower; otherwise, it will work faster. Thus, there is a need to create models that demonstrate the behavior of transistors in the case of deviating values. In other words, it is necessary to create models of transistors that will work quicker and slower. These models were acquired due to a deviation of the value of the threshold voltage variable. Thus, FF (fast-fast), SS (slow-slow) and TT (typical-typical) transistor models were developed.

During development of Monte Carlo models the above-mentioned method was used to find out the nominal values of statistical parameters. Afterwards a function called `agauss()` was used, to change statistical parameters values according to that function.

VI. THE TOOL DEVELOPMENT BASED ON THE ALGORITHM

During this research work a tool was developed with python programming language, which implements the algorithms described above, i.e. it automates the steps of the algorithms. The tool takes P, I and T functions as primary inputs and nominal values of parameters as optional inputs. If there are no given optional inputs, the tool applies default values to the parameters. As an output it develops the SPICE models according to the given inputs.

VII. ALGORITHM APPLICATION DURING DEVELOPMENT OF 14NM TRANSISTORS' SPICE MODELS

The full list of developed models with descriptions are the following:

- n08 – The SPICE model n type transistor.
- p08 - The SPICE model p type transistor.
- n08_hvt – The SPICE model n type transistor with high threshold.
- p08_hvt – The SPICE model p type transistor with high threshold.
- n08_lvt – The SPICE model n type transistor with low threshold.
- p08_lvt – The SPICE model p type transistor with low threshold.
- n08_slvt – The SPICE model n type transistor with superlow threshold.
- p08_slvt – The SPICE model p type transistor with superlow threshold.

In order to test the obtained model, the characteristics of transistors were compared with the characteristics introduced in different research papers.

All the deviations of the parameters were in the allowed range.

Fig. 4. and Fig. 5 provide the input characteristics of the "nfet" and "pfet" transistors, operating with the 0.8 V supply voltage.

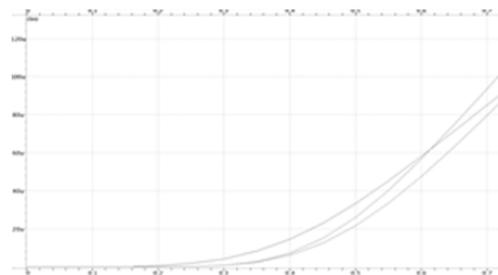


Fig. 5. Input characteristic of "nfet" transistor

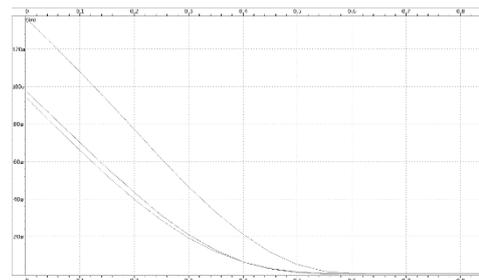


Fig. 6. Input characteristic of "pfet" transistor

VIII. CONCLUSION.

A new method has been developed that allows to get precise spice models of nanoscale transistors, which can be used in different studies and researches. A tool was developed with Python programming language which allows to automate the development of SPICE models of transistors. The tool was used to create SPICE models of transistors for 14nm. A research study of the obtained models was made. The results were compared with the characteristics of the foundry models of transistors in different research studies. Deviations were within the permissible limits. Also, Monte Carlo transistor models were developed, which will be used during the development of AOCV and POCV models.

REFERENCES

- [1] V. Melikyan, "Challenges and Solutions of IC Design Using FinFET Transistors", Proceedings of International Forum "Microelectronics". 2nd International Conference on Integrated Circuits and Microelectronic Modules, Crimea, pp. 343-347, September 26-30, 2016.
- [2] S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds., "INTRODUCTION TO MONTE CARLO SIMULATION", Simulation Conference, 2008. WSC 2008. Winter.
- [3] PrimeTime documentation, Synopsys Inc., 2017.
- [4] Arizona State University, MOSFET models, <http://ptm.asu.edu/modelcard/PTM-MG/modelfiles/hp/14nfet.pm>.
- [5] Goldman R., Bartleson K., Wood T., etc., "32/28nm Educational Design Kit: Capabilities, Deployment and Future, Microelectronics and Electronics (PrimeAsia)", IEEE Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia) - Visakhapatnam, India, 19-21 Dec. 2013.
- [6] Prateek M., Anish M., Niraj J., "FinFET Circuit Design", Springer 2010 - P. 23-53.
- [7] Tetsu Tanaka, "3D-IC technology and reliability challenges, Junction Technology (IWJT)", 2017 17th International Workshop on Junction Technology (IWJT) - Uji, Japan, 1-2 June 2017.
- [8] J. P. Duarte, S. Khandelwal, A. Medury, etc., "BSIM-CMG: Standard FinFET compact model for advanced circuit design", ESSCIRC Conference 2015 - 41st European Solid-State Circuits Conference (ESSCIRC), Graz, Austria, 14-18 Sept. 2015.

Comparison Of Grapheme-to-Phoneme Conversions For Spoken Document Retrieval

Dmitriy Prozorov
doctor of engineering sciences, professor at
Vyatka State University
prozorov.de@gmail.com

Alexandra Tatarinova
at
Vyatka State University
tatarinova.alexg@gmail.com

Abstract

The article contains analysis of spoken document retrieval techniques which apply word similarity based on phonemic transcriptions building or approximate string matching. Results are obtained on the collection of spoken documents with speech on Russian language. Grapheme-to-phoneme conversion methods based on a hidden Markov model and 1,2-order finite Markov chain is discussed on the article.

1. Introduction

Grapheme-to-phoneme conversion is a topical task that occurs in the field of speech technology. Examples are a preparation of a vocabulary for speech recognition and synthesis or an evaluation of word similarity for spoken document retrieval. Grapheme-to-phoneme conversion consists in building a phonemic transcription of a word which reflects a pronunciation of the word.

There are two approaches of phonemic transcription building [1]. Knowledge-based methods use a vocabulary or a set of linguistic rules which are hand-created by some experts [2-4]. Data-based methods are based on a transcribing algorithm trained on a vocabulary which contains presentations of words as sequences of graphemes and phonemes [1, 5]. The disadvantage of the first approach is finite of a vocabulary and need to manually compilation of a set of rules for building transcription. The second approach depends on quality of training data.

There are three groups of data-based methods [1]. Method of local classification sequentially defines a phoneme depending on a grapheme environment in a word. Method based on a pronunciation by analogy uses finding similar words or parts of words from the vocabulary. A transcription is built by analogy of

transcriptions of obtained words. The third method is based on a static model of pronunciation of words.

The article contains description of an application high-order Markov chain for building phonemic transcriptions of words. Accuracy of spoken document retrieval based on word similarity using phonemic transcriptions building or approximate string matching is analyzed for spoken documents contained speech on Russian language.

2. Task statement

Let an alphabet of phonemes be $\Phi = \{\varphi_j\}$ and an alphabet of graphemes (letters) be $C = \{c_i\}$, where $|\Phi| = N$ and $|C| = M$. And there is a training vocabulary Ψ which contains words. These words are presented as sequences of graphemes with phonemic transcriptions.

Need for a given word w to find a sequence of phonemes $\langle \varphi^0 \varphi^1 \dots \varphi^n \rangle$ which corresponds a pronunciation of the sequence of graphemes $\langle c^0 c^1 \dots c^n \rangle$

$$\varphi^0 \varphi^1 \dots \varphi^n = f(c^0 c^1 \dots c^n) = f(w), \quad (1)$$

where $c^k \in C$ and $\varphi^k \in \Phi$.

Grapheme-to-phoneme conversion can be applied computing phonemic similarity between two words. It is useful for Spoken Document Retrieval (SDR).

The SDR task is formulated as retrieving documents from a collection D of audio with speech which are relevant to a given text or spoken query Q . A query Q is a set of words $\{w_i\}$. Relevance score of document $d_i \in D$ in relation to the query Q is defined as

$$r_i = F(d_i, Q). \quad (2)$$

There are different information retrieval (IR) models which define a type of function (2) and a presentation technique of query and documents.

3. Markov chain of phonemic sequence

A phoneme occurrence in a word depends on previous phonemes in the word. Statistical dependence of a phoneme on a k -previous phoneme in a word transcription is decreasing. It can be characterized by values of a sum which is computed for a set of words

$$\frac{1}{N} \sum_{i=1..N} \max_{j=1..N} P(\varphi_i^l | \varphi_j^{l-k}), \quad (3)$$

where l is a position of a phoneme in a transcription and k is a shift of the position, $l = \bar{k}..n$ and n is a length of the transcription.

The histogram on the figure 1 shows values (3) according to k . The figure 1 demonstrates decrease of dependency between neighboring phonemes in a word according to a law which is approximation to exponential type.

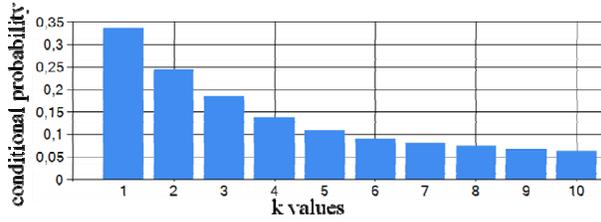


Fig. 1. Dependence of a phoneme on a k -previous phoneme

Theory of finite high-order Markov chain can be used building most probability phonemic sequence of a word according to Doob theorem [6].

Suppose that a phoneme in a transcription is defined t -previous phonemes as is shown on the figure 2. It allows to using t -order Markov chain for building a phonemic transcription [7].

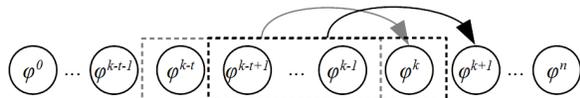


Fig. 2. Definition phoneme by t -previous phonemes

High-order Markov chain with N states can be transformed to simple Markov chain with N^t vectors of states [8].

A vector of state is

$$\vec{\varphi}_{ij..r}^k = (\varphi_i^k \varphi_j^{k-1} \dots \varphi_r^{k-t+1}) \quad (4)$$

and a vector of observable is

$$\vec{c}_{ij..r}^k = (c_i^k c_j^{k-1} \dots c_r^{k-t+1}). \quad (5)$$

A transition matrix for the Markov chain is

$$\Pi = [\pi_{ij..rq}]_{N^t \times N^t}. \quad (6)$$

where $\pi_{ijr} = P(\vec{\varphi}_{ij} | \vec{\varphi}_{jr}) = P(\varphi_i | \varphi_j \varphi_r)$.

Note that a transition graph of state vectors which is defined matrix (6) is a graph De Bruijn.

4. Phonemic transcription

Basically a phoneme is presented as a tuple of acoustic features. But a phoneme in a transcription is defined by rules of combinations of acoustic units in a word of a given language too.

Suppose a phoneme is described as distribution

$$P_{C|\varphi_j} \equiv P(C | \varphi_j), \quad (7)$$

where $P(C | \varphi_j)$ is a conditional probability of correspondence grapheme c_i to φ_j in transcriptions.

A posteriori probability of state vectors of finite Markov chain is presented in [10]

$$p^{ac}(\vec{\varphi}_{ij..r}^{\rightarrow k+1}) = c \cdot \exp\left(f(\vec{\varphi}_{ij..r}^{\rightarrow k+1})\right) \times \prod_{t=1}^N p^{ac}(\vec{\varphi}_{j_i..r,q_t}^{\rightarrow k}) \cdot P(\vec{\varphi}_{ij..r}^{\rightarrow k+1} | \vec{\varphi}_{j_i..r,q_t}^{\rightarrow k}), \quad (8)$$

where $f(\vec{\varphi}_{ij..r}^{\rightarrow k+1})$ is log-likelihood of a state vector $\vec{\varphi}_{ij..r}^{\rightarrow k+1}$ and c is normalization coefficient.

Thus a phoneme on k -position is defined according to a criterion of maximum a posteriori estimation

$$\varphi_i : (\varphi_i^k \varphi_j^{k-1} \dots \varphi_r^{k-t+1}) = \vec{\varphi}_{ij..r}^k \Rightarrow \Rightarrow i = i_t = \arg \max_{i=1..N} (u_{i,j_i..r_t}^k) \quad (9)$$

where

$$u_{ij..r}^{k+1} = \ln\left(p^{ac}(\vec{\varphi}_{ij..r}^{\rightarrow k+1})\right). \quad (10)$$

Example of building a phoneme transcription through 2-order Markov chain according to (9) is shown on the figure 3.

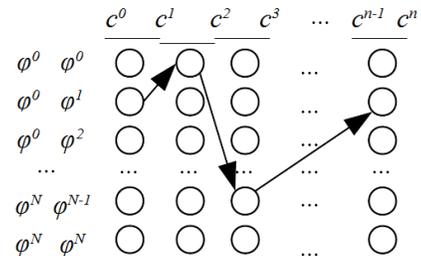


Fig. 3. Building phonemic transcription through 2-order Markov chain

A phoneme φ^k is defined by (9) using the word bigram $c^{k-1} c^k$ and the most suitable pair of alphabet phonemes.

5. Experiment

The experiment is contained in comparison of SDR accuracy according to word similarity measure. Values of accuracy is obtained on the collection of radio news [9] and the test query set [9]. Average duration of spoken documents of the collection is about 33 seconds. Used speech recognizers are CMU Sphinx [10, 11] and Yandex SpeechKit [12].

Spoken document retrieval systems and models for the experiment are described as

- Voice Digger [13] is a retrieval system of spoken documents which uses hidden Markov models (HMM) which are built on acoustic data.

- Lucene [14] is a fulltext retrieval system which realizes a fuzzy, a n-gram and others search methods. Base lucene method is a text search based on exact matching words. Fuzzy lucene method uses Levenshtein distance for matching words. N-gram lucene method presents n-gram searching through NgramTokenizer where N = 3, 4, 5.

- VSM* is a vector space model using the weighted cosine measure [15]. Weights of the measure are computed on values of word similarity. There are text and phonemic similarity. Text similarity based on the longest common substring or Levenshtein distance. Phonemic similarity uses phonemic transcriptions of words [7, 15] which are built HMM or high-order Markov chain.

SDR accuracy is evaluated as mean average precision (MAP). MAP values according to search method and speech recognizer are presented on the Table I.

TABLE I. MAP VALUES OF SPOKEN DOCUMENT RETRIEVAL

<i>Model or system</i>		<i>MAP</i>	
Voice Digger		0.6786	
<i>Recognizer</i>		<i>CMU Pocketsphinx</i>	<i>Yandex SpeechKit</i>
Lucene	Base	0.4442	0.7874
	Fuzzy	0.6514	0.8672
	Ngram	0.6429	0.8339
VSM*	Substring	0.6712	0.8929
	Levenshtein distance	0.6809	0.9015
	HMM transcribing	0.6817	0.8968
	Mark1 transcribing	0.6808	0.8907
	Mark2 transcribing	0.6821	0.8948

The best experimental result is retrieval technique based on VSM* with recognition through Yandex SpeechKit. Note that MAP values of retrieval based on VSM* and 1,2-order Markov chain with CMU Pocketsphinx recognizer are close MAP value of Voice Digger which is a commercial product.

Yandex SpeechKit is a web-service which allows access to speech recognition based on a neural network and using Yandex servers through HTTP protocol. CMU pocketsphinx is a library of speech recognition for a desktop. The library uses HMM acoustic model and n-gram language model. Accuracy of speech recognition for Russian language through Yandex SpeechKit is 80-90%. CMU Pocketsphinx amounts 40-80% accuracy of speech recognition for Russian language.

However, there are spoken documents which is hard for Yandex SpeechKit recognition. Accuracy of speech recognition for some documents is presented on the Table II.

TABLE II. EXAMPLE ACCURACY OF SPEECH RECOGNITION

<i>Document</i>	<i>CMU Pocketsphinx</i>	<i>Yandex SpeechKit</i>
0	54,70%	88,89%
1	35,14%	68,92%
2	20,74%	72,59%
3	28,50%	47,66%

Most CMU Pocketsphinx speech recognition errors is a distortion of words and most Yandex SpeechKit errors is a deletion of words. MAP values for hard documents which contain speech with record in a cave are presented on the Table III.

TABLE III. MAP VALUES OF HARD SPOKEN DOCUMENT RETRIEVAL

<i>Model or system</i>		<i>MAP</i>	
Voice Digger		0.4596	
<i>Recognizer</i>		<i>CMU Pocketsphinx</i>	<i>Yandex SpeechKit</i>
Lucene	Base	0.0588	0
	Fuzzy	0.2059	0.0441
	Ngram	0.4216	0.3284
VSM*	Substring	0.5699	0.3922
	Levenshtein distance	0.5294	0.4179
	HMM transcribing	0.5588	0.3897
	Mark1 transcribing	0.5907	0.4841
	Mark2 transcribing	0.5650	0.4179

The experiment result shows that using of grapheme-to-phoneme conversion based on 1,2-order Markov chain allows to increase MAP of retrieval for hard documents with continuous speech on Russian language.

7. Conclusion

Accuracy of SDR is highly dependent on quality of speech recognition. Speech on Russian language is hard to speech recognition. Using word similarity based on grapheme-to-phoneme conversion can be

increase accuracy of SDR for spoken documents which are recognized with low quality.

9. References

- [1] Bisani M., Ney H. Joint-sequence models for grapheme-to-phoneme conversion // SPECOM. 2008.
- [2] Kipyatkova I.S., Karpov A.A. The phonemic transcription module for the recognition system of spoken Russian speech // Artificial Intelligence, Donetsk, Ukraine, No. 4, 2008, S. 747-757. (in Russian)
- [3] Hunnicutt S. Grapheme-to-phoneme rules: A review // Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, QPSR 2-3. 1980. pp. 38-60.
- [4] Smirnov V.A., Gusev M.N., Farkhadov M.P. The function of the linguistic processor in the system of automatic analysis of unstructured speech information // Automation and modern technology. No. 8. 2013.P. 20-28. (in Russian)
- [5] Novak J., Minematsu N., Hirose K. WFST-based Grapheme-to-Phoneme Conversion: Open Source Tools for Alignment, Model-Building and Decoding // Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing. 2012. pp.45-49.
- [6] Doob J.L. Stochastic processes // New York: Wiley, 1990.
- [7] Prozorov D.E., Tatarinova A.G., Grapheme-to-phoneme conversion based on high-order Markov chain for spoken term detection by text query, IEEE East-West Design & Test Symposium (EWDTS), 2017, pp. 646-650
- [8] Prozorov D.E., Pletnev K.V., Yashina A.G., A posterior estimation of high-order Markov chain states, Information and space, 1(6), 2016, <http://openbooks.ifmo.ru/read/15422/15422.pdf> (in Russian)
- [9] Tatarinova A.G., Prozorov D.E. Building Test Speech Dataset on Russian Language for Spoken Document Retrieval Task, IEEE East-West Design & Test Symposium (EWDTS), 2018, pp. 884-887
- [10] CMU Sphinx, Open Source Toolkit For Speech Recognition. <http://cmusphinx.sourceforge.net>
- [11] Acoustic and Language model for Russian language (zero_ru_cont_8k_v3), [https://sourceforge.net/projects/cmusphinx/files/Acoustic and Language Models/Russian/](https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/Russian/)
- [12] Spoken speech recognizer Yandex SpeechKit: <https://tech.yandex.ru/speechkit/>
- [13] Keyword Spotting System Voice Digger // URL: <https://www.speechpro.ru/product/sistemy-upravleniya-kachestvom-i-avtomatizatsii/voice-digger>
- [14] Lucene .NET: <https://lucenenet.apache.org/>
- [15] Prozorov D.E., Yashina A.G., A weighted cosine measure of vector space model for spoken document retrieval, Information technology, №9, v. 21, 2015. – pp. 715-720. (in Russian)

Formalized Methods of Analysis and Synthesis of Electronic Document Management of Technical Documentation

Dilshod K. Baratov,
“Automation and telemechanics on railway transport” Department,
Ph. D., associate Professor.,
Tashkent Institute of railway engineers,
Tashkent, Uzbekistan
baratovdx@yandex.ru

Nazirjon M. Aripov,
“Automation and telemechanics on railway transport” Department,
DSc, Professor,
Tashkent Institute of railway engineers,
Tashkent, Uzbekistan
aripovnm@mail.ru

Davron Kh. Ruziev
“Automation and telemechanics on railway transport” Department,
assistant, Tashkent Institute of railway engineers
Tashkent, Uzbekistan
ruziyevddd@gmail.com

Abstract—The article discussed the problem of the synthesis of algorithmic display of systems of railway automation and remote control based on the process of accounting and control of automation devices and telemechanics. This process is described using a formalized description language. Considering the verification of some properties of control algorithms and their transformation in order to optimize the structure of the process of accounting and control of automation and remote control devices, the language of logic circuits of algorithms was chosen. As a formalized language, logical schemes of algorithms were chosen. As a graphical representation of the process of accounting and control of automation and telemechanics devices, a graph-scheme of algorithms is used. The condition of functioning of process of the account and control of devices of automation and telemechanics is described by means of matrix schemes of algorithms.

The proposed formal scheme of the process of accounting for and control of devices of automatics and telemechanics using the logical schemes of algorithms provides the formalisation of the procedure for the transition to automated technology. The formalized scheme makes it possible to conduct high-quality and accelerated operational research electronic document management of accounting and control of automation and telemechanics devices in automation and telemechanics system's in railway transport.

Keywords—*accounting and control of railway automation and remote control devices, logic schemes of algorithms, state minimization method, technical documentation.*

I. INTRODUCTION

In connection with the need for broad modernization, reconstruction and replacement of railway automation devices, an important task is to improve the quality of the process of control and accounting of railway automation and remote control devices (CARCD). The existing process technology CARCD does not ensure the adoption of quick and effective decisions.

Existing software mainly focuses on the design of printed circuit boards, programmable logic matrices, analog, digital and mixed analog-digital devices, electrical circuits of power electrical equipment of engineering companies, synthesis of devices programmable by logic and analog filters, automation of design work when creating electrical control systems database of contact equipment and programmable controllers. But for the task of automating the process of technical documentation management, and process control and accounting of railway automation and remote control devices the existing software is not suitable.

To solve the tasks of organizing electronic document management of technical documentation, and the process of monitoring and recording devices for railway automation and remote control, it is necessary to develop specialized software packages that take into account the specifics of all stages of the life cycle of alarm systems, centralization and blocking.

To solve this problem in this work it is proposed to create a model of CARCD as an electronic document management of technical documentation for signalization, centralization and blocking devices. In this connection, a survey was made of the real processes of creating, verifying and using technical documentation in automation and remote control systems. This made it possible to identify document flow scenarios and protocols for technical document properties [1,2].

The most effective solution to the problems of automating the CARCD process can be achieved by formalizing the processes of the CARCD and applying mathematical methods to optimize the coordination of interaction.

Formal methods for displaying processes are used to analyze the properties of an object by formal models. Thus, the properties of an object must be formalized within the framework of a certain mathematical model. Accordingly, for the application of formal methods for describing the CARCD process and for determining the composition and properties of the standard means for describing CARCD, it is necessary to develop a formalized scheme for describing objects involved in the CARCD process.

A formalized scheme is a form that determines the composition and type of source data and should provide the possibility of describing an object in a volume sufficient to automate the process of CARCD.

A formalized scheme should be universal, i.e. sufficiently generalized to describe a wide class of objects (specifications, orders, applications) and at the same time ensure the simplicity of the procedures for linking to a specific object. For this, a formalized scheme should include a set of tools for representing the elements, structure and algorithms of the system, functional and static dependencies between parameters.

There are a number of methods for identifying algorithms for the operation of complex systems, namely: the method of simplifying the work; compilation of structural information-time diagrams, flowcharts and organigrams [3, 4]. The essence of the indicated methods consists in the operation recording and analysis of the process under study. The common shortcomings of these methods from the point of view of the study of CARCD on the basis of the proposed formalized schemes are: a limited set of symbols for operations; complexity, and for a number of methods and the impossibility of displaying parallel processes, the complexity of filling out survey forms. In this regard, in this work, for constructing an algorithmic mapping CARCD, it is proposed to use the languages of direct description of discrete processes, which include Petri nets [5], logic schemes of algorithms LSA [6, 7], logic schemes of requirements [8], parallel logic schemes of algorithms PLSA [9,10]. The necessity of combining CARCD algorithms requires providing the possibility of formalized transformation of algorithms. The need to meet these requirements leads to the choice of language LSA, which are shown in [11-13].

LSA is used as a language for setting algorithms for the operation of software control devices [14].

II. FORMALIZED SCHEME OF TECHNICAL DOCUMENTATION

The LSA searches for an algorithm for processing some initial information, that is, in the selection of individual operations, or acts of the algorithm, and the search for the order of their execution. Each such act (operation) in the LSA is matched with an operator, denoted by capital Latin letters $A, B, C \dots$. Different operators may be denoted by different letters or by the same letter, but with different indices: $A1, A2, \dots, B1, B2 \dots$. If the operator depends on parameters, then these parameters can be set as indices $Ai, Aij, Aijk \dots$ or in brackets: $A(i), A(ij), A(ijk) \dots$. Operators with different parameters, they perform actions on different parts of the original or intermediate data, i.e., on different parts of the processed information.

The process of accounting and control of railway automation and remote control devices is described as follows:

$$p_g \in P, g = \overline{1, G} \quad (1)$$

where p_g – is the device (device), the set G forms a P set of devices. Also determined by the parameters of the device:

$$h_{g,m} \in H_g, m = \overline{1, M} \quad (2)$$

$P_{g,m}$ – m the instrument parameter p , the instrument parameter set M forms the sets H_p of all the considered parameters of the instrument g (each m parameter is entered in its instrument position).

Definition 1. The set of operations and checks of logical conditions performed in a specific sequence in the CARCDprocess is an algorithm A_g .

Definition 2. The operation O_p is an elementary action to account for and control devices from the set of S . All operations performed in the process of accounting and control $s_g \in S$, form a set of $O = \{o_p\}, p = \overline{1, P}$. The index of the operation $O_p, p = \overline{1, P}$ specifies the number of the participant and the algorithm, as well as its individual number in the sequence of entries.

In this work the symbolism of the record of private algorithms of CARCD is introduced in the language of parallel logic schemes of algorithms (PLSA) with regard to the generalized formalized scheme [15]. The main elements are operators corresponding to operations O_p , logical conditions $\alpha_k, k = \overline{1, K}$, marked with arrows $\alpha_k \uparrow^p, p = \overline{1, P}$, where p is the index of the arrow. The transition with a false value α_k is carried out to the element of the PLSA, marked with an arrow with the same index \downarrow^p .

To account for the quality CARCD, additionally introduced the following types of operations:

j – j -th line;

α_t – waiting logical condition: $\alpha_t = 1$ if $f(t) \geq T_H$

\downarrow^k – the end of algorithm

k – operations, that determine the quality of CARCD (control parameters);

α – probabilistic logical conditions that depend on quality CARCD.

The set $A = \{\alpha_k\}, k = 2$ includes the probabilistic logical conditions of the form

$$\alpha_k = \begin{cases} 1 - \text{positive result;} \\ 0 - \text{otherwise;} \end{cases}$$

The sequence of execution of operators in the LSA is determined by the order in which they are written. For example, $A11A12A13$ means that the operator $A11$ is first executed, then $A12$, and then $A13$. The order of execution of operators in the LSA can be strictly fixed — a linear algorithm — or depending on certain conditions — a branched algorithm. In the latter case, the LSA uses logical conditions denoted by small Latin letters, $p, q, r \dots$. Like

operators, different logical conditions (LC) are denoted by different letters or by the same letter, but with different indices.

Logical conditions may depend on several variables. Logical conditions depending on the values of the function n variables are denoted by

$$\alpha [f(\alpha_1, \alpha_2, \dots, \alpha_n)] \quad (3)$$

It is considered that logical conditions can take only two values: the condition being checked is satisfied ($\alpha_i=1$) or not ($\alpha_i=0$). Depending on the value of the currently checked LC, the further order of execution of the operators and LC is determined.

Often among logical conditions it is advisable to select those that always take a zero (false) value, that is, identically *false logical conditions*. Identical logical conditions do not require verification. They are denoted by ω . Operators and LCs are basic, and identically false logical conditions are auxiliary members of the logical scheme of the algorithm.

Each LC has an arrow. The beginning of the i -th arrow (denoted by \uparrow^i) is to the right of the logical condition, and its end (denoted by \downarrow^i) is to the left of the LSA member that must be met if the LC takes a zero value.

LSA are called expressions made up of operators following each other and LC, as well as numbered arrows arranged in a certain way. The logic scheme of the algorithm is some way of describing the algorithm for solving the problem [16,17].

Description of the algorithm using logic circuits is the first step in the formalization of the algorithm. This stage is preceded by a meaningful description of the algorithm. The logic scheme of the algorithm allows both formal and informative equivalent transformations.

III. METHOD OF MINIMIZATION OF LSA OF TECHNICAL DOCUMENTATION

LSA A_U will be represented using the function LC:

$$A_U = f_1(\alpha_1\alpha_2\alpha_3)Z_1 \vee f_2(\alpha_1\alpha_2\alpha_3)Z_2 \vee \dots \vee f_8(\alpha_1\alpha_2\alpha_3)Z_8 \quad (5)$$

where, Z_1, Z_2, \dots, Z_8 the value of A_U after LC function.

In summary

$$A_U = \sum_{m=1}^M f_m(\alpha_1\alpha_2\alpha_3)Z_m \quad (6)$$

- 1) if $\alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0$, that
 $k2 A24 A31 A33 A34 k4 k5 A35$;
- 2) if $\alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 1$ that

As a result of the analysis of the processes included in CARCD at all levels, the LSA was received in the form:

$$A_U = A11 \downarrow^3 A12 k1 A21 A22 \alpha_1 \uparrow^1 A23 \times \alpha_2 \uparrow^2 k1 A13 \omega \uparrow^3 \downarrow^1 k2 \downarrow^2 A24 \alpha_3 \uparrow^4 \times k3 A14 k1 A41 A42 k6 A43 k7 A51 A52 \times A53 k7 A61 A62 A63 k1 A32 \omega \uparrow^5 \downarrow^4 \times A31 \downarrow^5 A33 A34 k4 k5 A35 \downarrow^k \quad (4)$$

where $\alpha_1, \alpha_2, \alpha_3$ – logical conditions, the probability of fulfillment which depends on the current value of the quality indicator CARCD.

Logic scheme determines the order of execution of operators depending on the value of the LC included in it.

The algorithm begins with the execution of the leftmost operator of the scheme. After the operators of the scheme A_U are executed, it is determined which operator of the scheme should follow after it. After the operator $A11$, the operator of the scheme that is directly to its right ($A12$) must be executed. After the logical condition α_1 , two cases are possible: if the condition being checked is executed, then operator $A23$ on the right must be executed; if it is violated, then the operator $A24$ is executed, to which an arrow leads, starting after this condition

The algorithm ends when the last of the executing operator $A63$ contains an indication of the termination of the algorithm.

Each elementary operation of the A_U algorithm, in turn, is represented by a lower level algorithm in its alphabet of operators. By this is achieved the construction of a hierarchical structure of the description of the CARCD.

where, $m = \overline{1, M}$.

Thus, the distribution of values of LC in a logical scheme determines the order execution, included in this scheme.

Since each of the LC can take only two values - 0 and 1, the maximum number of unique sets of LC, and, therefore, the number of rows in the truth table, can be determined by the formula: $N = 2^n$, where 2 – base of the number system (all LCs can take only one of two possible values); n – LC number.

In the logical scheme (4), the order of execution of operators depending on the values of the LC is as follows:

- $k2 A24 k3 A14 k1 A41 A42 k6 A43 k7 A51 A52 \times A53 k7 A61 A62 A63 k1 A32 A33 A34 k4 k5 A35$;
- 3) if $\alpha_1 = 0, \alpha_2 = 1, \alpha_3 = 0$ that
 $k2 A24 A31 A33 A34 k4 k5 A35$;
- 4) if $\alpha_1 = 0, \alpha_2 = 1, \alpha_3 = 1$ that

$k2 A24k3A14k1 A41A42k6A43k7A51A52 \times$
 $A53k7A61A62A63k1A32A33A34k4k5A35$;

5) if $\alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 0$ that

$A23A24A31A33A34k4k5A35$

6) if $\alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 1$ that

$A23A24k3A14k1 A41A42k6A43k7A51A52 \times$
 $A53k7A61A62A63k1A32A33A34k4k5A35$

7) if $\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 0$ that

$A23k1A13A12k1A22$;

8) if $\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1$ that

$A23k1A13A12k1A22$.

The truth table of A_U is presented in table 1.

From the possible values of A_U are chosen the similar LSA. It is not difficult to notice, that the values of 0, 2 and

4 sets (with the difference of the operator $A23$ and $k2$), 1, 3 and 5 sets (with the difference of the operator $A23$ and $k2$), 6 and 7 sets are similar. The same parts of the LSA are denoted as follows.

TABLE I. A_U TRUTH TABLE FOR THREE LCS

α_3	α_2	α_1	A_U algorithm value
0	0	0	$k2 A24 A31 A33 A34 k4 k5 A35$
0	0	1	$k2 A24 k3 A14 k1 A41 A42 k6 A43 k7 A51 A52 A53 k7 A61 A62 A63 k1 A32 A33 A34 k4 k5 A35$
0	1	0	$k2 A24 A31 A33 A34 k4 k5 A35$
0	1	1	$k2 A24 k3 A14 k1 A41 A42 k6 A43 k7 A51 A52 A53 k7 A61 A62 A63 k1 A32 A33 A34 k4 k5 A35$
1	0	0	$A23 A24 A31 A33 A34 k4 k5 A35$
1	0	1	$A23 A24 k3 A14 k1 A41 A42 k6 A43 k7 A51 A52 A53 k7 A61 A62 A63 k1 A32 A33 A34 k4 k5 A35$
1	1	0	$A23 k1 A13 A12 k1 A22$
1	1	1	$A23 k1 A13 A12 k1 A22$

TABLE II. THE SAME PARTS OF THE LSA A_U

LSA	Selection element	Name of the common part of the LSA
$A24 A31 A33 A34 k4 k5 A35$	$k2$	L1
$k2 A24 k3 A14 k1 A41 A42 k6 A43 k7 A51 A52 A53 k7 A61 A62 A63 k1 A32 A33 A34 k4 k5 A35$	$k2$	L2
$k2 A24 A31 A33 A34 k4 k5 A35$	$k2$	L1
$k2 A24 k3 A14 k1 A41 A42 k6 A43 k7 A51 A52 A53 k7 A61 A62 A63 k1 A32 A33 A34 k4 k5 A35$	$k2$	L2
$A23 A24 A31 A33 A34 k4 k5 A35$	$A23$	L1
$A23 A24 k3 A14 k1 A41 A42 k6 A43 k7 A51 A52 A53 k7 A61 A62 A63 k1 A32 A33 A34 k4 k5 A35$	$A23$	L2
$A23 k1 A13 A12 k1 A22$	-	L3
$A23 k1 A13 A12 k1 A22$	-	L3

Being constructed a transition table for this LSA

TABLE III.
TRANSITION TABLE LSA A_U

$\alpha_1\alpha_2$		00	01	10	11
α_3	0	k2L1	k2L1	A23L1	L3
	1	k2L2	k2L2	A23L2	L3

With the graphical method, each set of values of an LC corresponds to a certain point of n -dimensional space. The coordinates of the vertices of the n -dimensional cube correspond to the sets of values of the LC, and their designations are assigned the values of A_U on these sets. Since each of the LC can take only two values: 0 and 1, each edge connecting two adjacent vertices, the sets of which differ by one variable, has a unit length. Therefore, a n -dimensional cube is called a unit cube.

The number of vertices of a n -dimensional cube is equal to the number of rows in the truth table, and the number of coordinate axes is equal to the number of n LC.

3-x dimensional cube corresponding to the LC given earlier by the truth table (table. 1), shown in Fig.1. The top of the cube and the table cell.3, the contents of which describe the same set of variables (№6), are dotted. Similarly, the vertices of the cube are matched to the remaining cells of the truth table.

Using the coordinate method, the LSA A_U is given as a state coordinate map, called a Carnot map. The total number of cells in the Carnot map corresponds to the number of sets of the algorithm A_U .

Designations in Carnot map for LSA A_U :

$$\alpha_i \rightarrow (\alpha_i = 1), \bar{\alpha}_i \rightarrow (\alpha_i = 0)$$

In brackets inside the cells are the numbers of the corresponding sets from the truth table (Table 3).

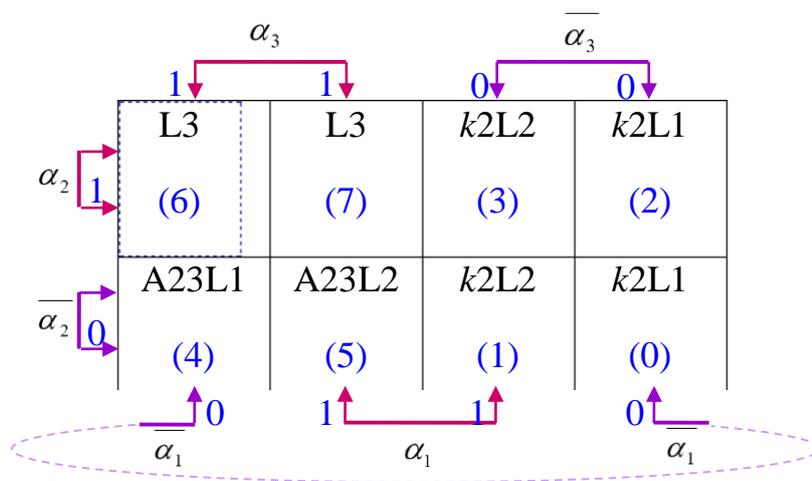


Fig. 1. Carnot map for LSA A_U

The task of minimizing the LSA A_U is to find the minimum set of terms with the smallest number of LC, allowing to get all the output value.

In (5) Z_1, Z_2, \dots, Z_9 is not a variable. If accept the condition that the symbol Z_m should always remain at the end of the addendum, and the inversion operation should not be applied to it, then can be formally operate with it as with the variable.

All cells containing the same LSA value are combined into closed areas, each area must represent a rectangle with the number of cells 2, 4, 8. The areas can intersect, and the same cells can belong to different areas. Neighboring cells are not only cells that are adjacent horizontally and vertically, but also cells that are located on opposite borders of the map.

When covering cells with closed areas, one should strive for the minimum number of areas, each of which would contain as many cells as possible. Each member of the function of the LC is only one of those LCs that have one value for the corresponding area. If an LC for one cell of a region has one value, and for another cell of this region, another, it is not present in the corresponding member of the function of the LC.

To obtain the minimum form of the function in this work, it is proposed to cover the cells with the same LSA values with closed areas and to take the LC values with the same value within the corresponding areas when writing the members.

So, from Figure 2 is obtained the function of conditions $f(\alpha_1\alpha_2\alpha_3)$ for LSA A_U :

$$f = \alpha_2\alpha_3 \vee \alpha_1\bar{\alpha}_3 \vee \bar{\alpha}_1\bar{\alpha}_3 \quad (8)$$

Having executed minimization, is received:

$$f = \alpha_2\alpha_3 \vee \bar{\alpha}_3(\alpha_1 \vee \bar{\alpha}_1) = \alpha_2\alpha_3 \vee \bar{\alpha}_3 \quad (9)$$

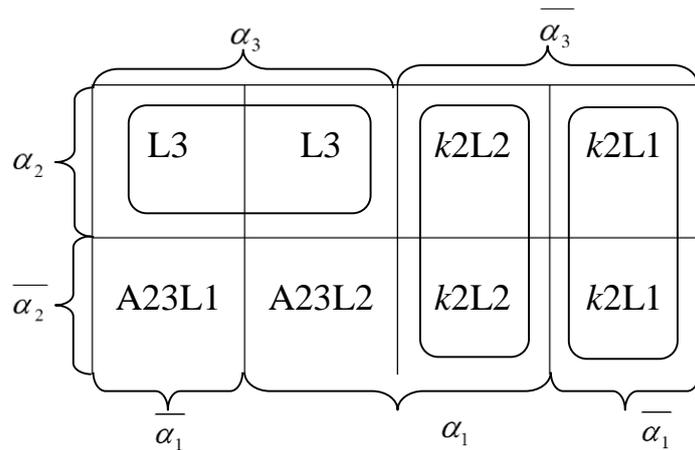


Fig. 2. LSA Truth Table Sets A_U

IV. MATRIX DIAGRAM ALGORITHMS OF THE PROCESS OF ACCOUNTING FOR AND CONTROL OF DEVICES OF AUTOMATICS AND TELEMCHANICS

The matrix scheme of the algorithm (MSA) is a square matrix (table 3), each row and each column of which is associated with the operator.

$$M = \begin{matrix} & A_1 & A_2 & \dots & A_k \\ A_0 & a_{0,1} & a_{0,2} & \dots & a_{0,k} \\ A_1 & a_{11} & a_{12} & \dots & a_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ A_{k-1} & a_{k-1,1} & a_{k-1,2} & \dots & a_{k-1,k} \end{matrix}$$

The matrix element CARCD a_{ij} is the logical function of logical conditions. In this case, the operator A_g associated with the g -th column of the matrix is executed after the operator A_q associated with the q -th row, if the logical function

$$a_{qg} = a_{qg}(p_1, \dots, p_m) = 1$$

Logical functions MSA CARCD have the following two properties:

the consistency of the algorithm $a_{qg} a_{qj} = 0, g \neq j$;

completeness of the algorithm $\bigvee_{g=1}^l a_{qg} = 1$.

The product of two different functions of the same MSA row is always 0. This is the first condition necessary in order that, after the operator A_q could not run more than one operator. The second condition implies that at least one statement must always be executed after the operator A_q . Thus, only one operator is always executed after the operator A_q .

Prepare MSA to LSA the process of replacing the unit distance signaling and communication of the First, the MSA string associated with the V_0 operator is filled in. For this are

all functions a_{Aq} . After the V_0 operator, the V_{711} operator is always executed, i.e. $a_{V_0 V_{711}} = 1$. The function a_{Aq} will accept up to V_{718} . In this step, if $\alpha_{711} = 1$, then after V_{714} , you must run the V_{719} operator. Therefore, $a_{V_{718} V_{719}} = \alpha_{711}$. If $\alpha_{711} = 0$, then after V_{718} it is necessary to execute the V_{719} operator. Thus, $a_{V_{718} V_{719}} = \alpha_{711}$. Element $a_{V_{718} V_{719}}$ is equal to zero, since for any sets of values of LC after the operator V_{718} cannot be directly performed by the operator V_{719} .

Similarly to form other logic functions a_{Aq} , is compiled by the MSA, the appropriate LSA algorithm A71 (4).

$$\begin{aligned}
 & V_0 V_{711} V_{712} V_{713} V_{714} \downarrow^{717} V_{715} V_{716} V_{717} \downarrow^{715} V_{718} \times \\
 & \times \alpha_{711} \uparrow^{711} V_{7110} V_{7111} \overline{\alpha_{712}} \uparrow^{712} \downarrow^{711} V_{719} \omega \uparrow^{715} \times \\
 & \times \downarrow^{712} \downarrow^{714} V_{7115} \alpha_{713} \uparrow^{713} V_{7116} \alpha_{714} \uparrow^{714} V_{7117} \times \\
 & \times \omega \uparrow^{716} \downarrow^{713} V_{7118} \omega \uparrow^{717} \downarrow^{712} V_{7112} V_{7113} V_{7114} V_k
 \end{aligned} \quad (10)$$

For convenience, used graph-schemes of algorithms [18-21], which give a more visual representation of the algorithm A71. From Fig.3 and LSA A71 their mutual connection is visible, and on the graph diagram operators are specified by rectangles, and logical conditions — circles. The initial operator (V_0 operator) has no input arrows, the final operator (V_k operator) has no output arrows. The unit value of the LC in the diagram graph corresponds to an arrow marked with a + sign, and the zero value to an arrow marked with a - sign.

Logical functions MSA CARCD possess the properties of consistency algorithm and the completeness of the algorithm.

Developed matrix diagram algorithms and graph-scheme of algorithm process CARCD clearly shows the correlation between Boolean terms and operators, the logical schemes of algorithms, i.e. the reciprocal relationship between participants, processes and the conditions of technical documentation.

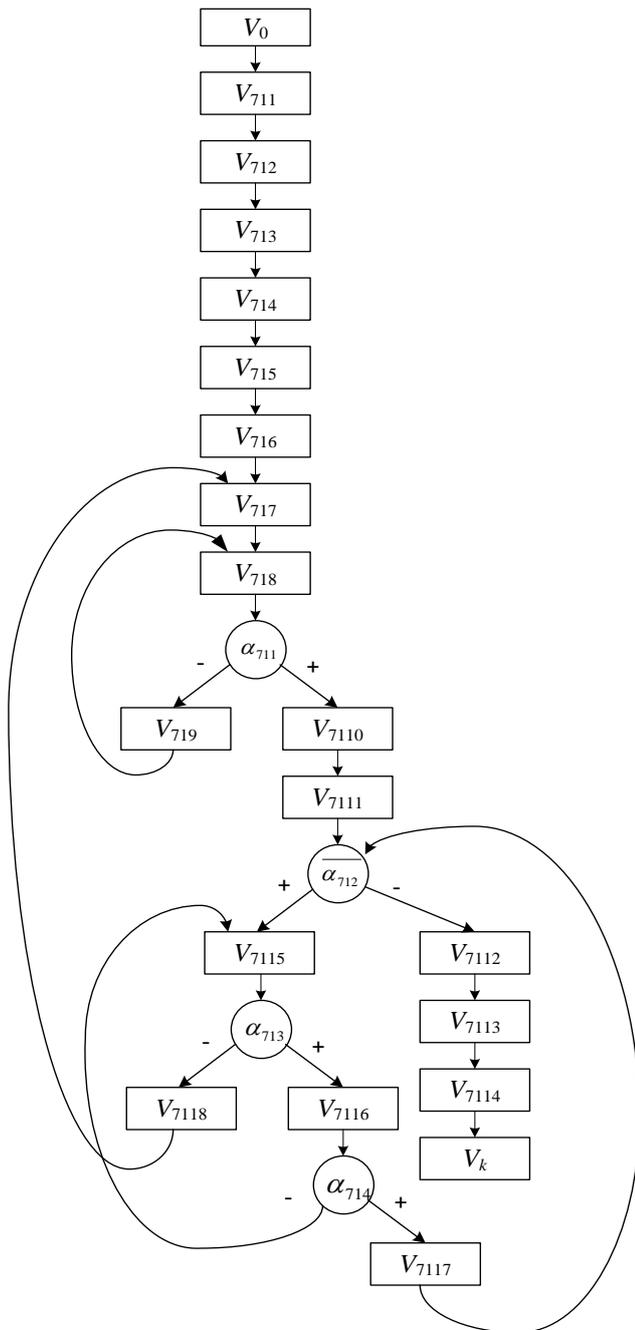


Fig. 3. Graph diagram of the LSA A71

The results of the research in the form of the theory and methods of organizing electronic document management of technical documentation are implemented in the software “Automated system of accounting and control of railway automation and telemechanics devices” (ASU-KZHAT), which was tested in Tashkent signaling and communication distance ShCh-1 JSC “Uzbekistan temir yo‘llari” specialists of the Department of automation and remote control at the railway transport of the Tashkent institute of railway engineers.

V. CONCLUSION

A formalized scheme provides sufficient flexibility to describe CARCD, because it is based on the algorithmic

representation of the system. In accordance with this method of formalization is aimed primarily at the identification and description of algorithms CARCD.

The use of the language of logical schemes of algorithms to identify and describe the processes of CARCD in railway transport has allowed to develop a new survey methodology aimed at identifying the structural-algorithmic and parametric display of the processing system.

The proposed method of minimizing the LSA allows you to get all the output value when finding the minimum set of members with the least number of LC.

Developed logical function MSA CARCD possess the properties of consistency and completeness of the algorithm. Developed by the matrix circuit and the graph-scheme of algorithm process CARCD clearly shows the correlation between Boolean terms and operators, the logical schemes of algorithms, i.e. the reciprocal relationship between participants, processes and the conditions of technical documentation.

The synthesized hierarchical structure of the model of electronic document management of technical documentation, where operators matrices of the highest level are represented as algorithms in a matrix of lower rank in their alphabet operations. The developed matrix model provides operational studies of electronic document management of technical documentation systems in the economy of automation and telemechanics in railway transport.

Using the language of logic circuits of algorithms for identifying and describing CARCD processes on railway transport allowed us to develop a new survey methodology aimed at identifying the structural-algorithmic and parametric mapping of the processing system.

The proposed formalized scheme of the CARCD process with the use of logic circuits of algorithms ensures the formalization of procedures for the transition to automated technology. The formalized scheme makes it possible to carry out high-quality and accelerated operational research of the electronic document flow CARCD in the systems of automation and telemechanics on the railway transport.

The proposed LSA minimization method allows you to get the entire output value when finding the minimum set of terms with the lowest number of LCs.

The developed MCA logic functions of CARCD have the consistency and completeness properties of the algorithm.

The developed matrix scheme and the graph-diagram of the algorithms of the CARCD process vividly show the interrelation between logical conditions and the operators of logical algorithms, i.e. mutual communication between participants, processes and technical documentation states.

The hierarchical structure of the electronic document management of technical documentation model is synthesized, in which the operators of higher-level matrices are represented as algorithms in lower-level matrices in their alphabet of operations. The developed matrix model provides operational research of electronic document management of technical documentation systems in the automation and telemechanics business of railway transport.

TABLE IV. MSA A71

	V_{711}	V_{712}	V_{713}	V_{714}	V_{715}	V_{716}	V_{717}	V_{718}	V_{719}	V_{710}	V_{7111}	V_{7112}	V_{7113}	V_{7114}	V_{7115}	V_{7116}	V_{7117}	V_{7118}	V_k
V_0	1																		
V_{711}		1																	
V_{712}			1																
V_{713}				1															
V_{714}					1														
V_{715}						1													
V_{716}							1												
V_{717}								1											
V_{718}									$\overline{\alpha_{711}}$	α_{711}									
V_{719}								1											
V_{7110}										1									
V_{7111}											α_{712}				$\overline{\alpha_{712}}$				
V_{7112}												1							
V_{7113}													1						
V_{7114}																			1
V_{7115}																α_{713}		$\overline{\alpha_{713}}$	
V_{7116}															$\overline{\alpha_{714}}$		α_{714}		
V_{7117}											α_{712}				α_{712}				
V_{7118}					1														

REFERENCES

[1] Igarashi S. On the logical schemes of algorithms, Information Processing in Japan, 1963, Vol. 3, pp. 12-18.

[2] Baratov D. X. The issues of creating a formalized model of the technical documentation, International Journal of InterScience, 2017, No.4 (1), pp. 22-23.

[3] Aripov N. M., Baratov D. K., Tokhtamysova A. B. Formalization of electronic technical document management of railway automatics and telemechanics, Bulletin of the Kazakh Academy of Transport and Communications, 2016, No..3, pp. 175-180.

[4] McAllester D. A logical algorithm for ML type inference, International Conference on Rewriting Techniques and Applications, Springer, Berlin, Heidelberg, 2003, pp. 436-451.

[5] Baccarne R. Hello CRIS, can a Library Software solution help you? //Procedia computer science. – 2019. – T. 146. – C. 208-219.

[6] Pilorget L., Schell T. IT Management: The art of managing IT based on a solid framework leveraging the company’s political ecosystem. – Springer, 2018.

[7] Baez J. C., Foley J., Moeller J. Network Models from Petri Nets with Catalysts //arXiv preprint arXiv:1904.03550. – 2019.

[8] Omar L. et al. A thematic analysis of technical documents: The collection and formalization of information relating to the needs of persons with disabilities, International Journal of Cognitive Research in Science, Engineering and Education, 2017, Vol. 5, No. 2.

[9] Furth S., Baumeister J. Semantification of large corpora of technical documentation, Enterprise Big Data Engineering, Analytics, and Management, IGI Global, 2016, pp. 171-200.

[10] Necker M., Necker M. C. A graph-based scheme for distributed interference coordination in cellular OFDMA networks, VTC Spring 2008-IEEE Vehicular Technology Conference, IEEE, 2008, pp. 713-718.

[11] Mendoza A. R. et al. Electronic document management system implementing i things (iot) internet of, International Journal of Advanced Research in Computer Science, 2019, VOL. 10, No. 2.

[12] Hirota M. et al. Document management system including image processing server and document management server, and document management server: Vol. 16156402 USA, 2019.

[13] Aguilar A., Lozoya C., Orona L. M. A hamming distance and fuzzy logic-based algorithm for P2P content distribution in enterprise networks, Peer-to-Peer Networking and Applications, 2019, pp. 1-13.

[14] Zakoldaev D. A. et al. Computer-aided design of technical documentation on the digital product models of Industry 4.0, IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2019, VOL. 483, No. 1, pp. 012069.

[15] Talamo C., Atta N. Management of FM-related Information, Invitations to Tender for Facility Management Services, Springer, Cham, 2019, pp. 93-131.

[16] Brayton R. K. et al. Logic minimization algorithms for VLSI synthesis, Springer Science & Business Media, 1984, Vol. 2.

[17] Mao S., Rosenfeld A., Kanungo T. Document structure analysis algorithms: a literature survey, Document Recognition and Retrieval

- X, International Society for Optics and Photonics, 2003, VOL. 5010, pp. 197-208.
- [18] Coudert O. Two-level logic minimization: an overview, *Integration*, 1994, VOL. 17, No. 2, pp. 97-140.
- [19] Baratov D. K., Aripov N. M. Formalization of electronic technical document management of railway automatics and telemechanics, *Europäische Fachhochschule*, 2016, No. 8, pp. 33-35.
- [20] Aripov N., Baratov D. Features of Construction of Systems of Railway Automatics and Telemechanics at the Organization of High-Speed Traffic in the Republic of Uzbekistan, *Procedia Engineering*, 2016, VOL. 134, pp. 175-180.
- [21] Baratov D. K., Aripov N. M. Formalization of electronic technical document management of railway automatics and telemechanics, *Europäische Fachhochschule*, 2016, No. 8, pp. 33-35.

Non-Invasive System for Determining the Level of Iron in the Blood

Andrey Azarov
Dept. of Electronics Engineering
Sevastopol State University
Sevastopol, Russia
azarov@ieec.org

Elena Shirokova
Dept. of Radio Engineering and
Telecommunication
Sevastopol State University
Sevastopol, Russia
shirokova@ieec.org

Igor Shirokov
Dept. of Electronics Engineering
Sevastopol State University
Sevastopol, Russia
shirokov@ieec.org

Abstract — In this article the technique of non-invasive determination of a iron level in a blood and, therefore, hemoglobin is proposed. The system operates on the basis of changes in the optical transparency of tissue under the influence of an external magnetic field, based on the fact that the tissue is saturated with blood containing the iron. Since the change in transparency is extremely small, a number of measures have been taken to highlight the useful signal, namely, at first, a constant component determined by the average illumination of the receiver is excluded, and then the signal is detected through the implementation of synchronous detection, which allows recognizing the signal at a signal-to-noise ratio of -100 dB.

Keywords— hemoglobin, anemia, iron in the blood, magnetic field, optical radiation)

I. INTRODUCTION

The situation related to the reduced level of hemoglobin in the blood of the population is relevant, because of the difficulties associated with the effects of this pathology. Hemoglobin is a respiratory iron-containing pigment of human blood. The main component of hemoglobin is iron, and in the role of the nitrogen base is a globin-protein, as it is shown in Fig. 1.

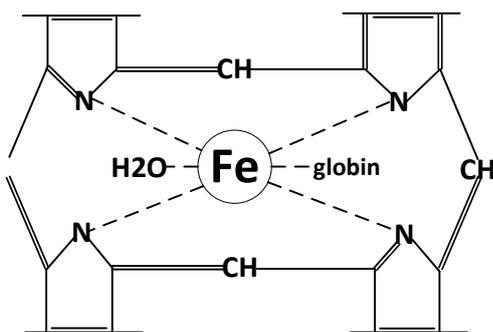


Fig. 1. Chemical structure of hemoglobin based on an iron molecule.

Reduced hemoglobin levels in humans cause such a disease as anemia. Depending on the degree of the disease [1, 2], symptoms may vary, but ones are appeared in following: people are usually weak, quickly tired, dyspnea and paleness of the skin and mucous membranes are appeared. Deep anemia may also cause frequent dizziness, tinnitus and visual impairment. According to the World Health Organization, there are more than 1.5 billion people in the world suffering from various degrees of this pathology.

One of the most important tasks of solving this problem is the difficult and time-consuming analysis. There are a number of methods and systems to determine the level of hemoglobin [3-8]:

- Calorimetric methods are based on the ability of blood hemoglobin to produce colored complexes during chemical reactions;
- Semiquantitative method is a method of relative determination of hemoglobin by the specific weight of blood;
- Gasometric methods are based on the use of the property of hemoglobin to add oxygen or carbon monoxide in strictly defined amounts;
- Calculation of hemoglobin concentration by the amount of iron in the blood sample.

These methods have proved to be good, but have a number of significant drawbacks due primarily to the invasiveness of the methods.

In modern literature there are described also a number of studies on the determination of hemoglobin levels in the blood. For example, in [9] there are described studies in which children suffering from moderate anemia were observed, and suggests a variant of complex analysis of children's body parameters for more accurate determination of the patient's condition. However, the method is also based on invasive systems, except for the analysis of pathology symptoms, which is not a reliable factor.

In the paper [10] the technique of noninvasive determination of the hemoglobin level by means of the use of pulse oximetry is proposed. This technique implies the presence of a system similar to the transceiver, which is quite difficult. It is also necessary to take into account the fact that the method is designed for determining the saturation of the oxygen, and therefore the determination of hemoglobin or iron is an indirect factor, which is unforgivable for medicine.

The main problems with this type of system include the duration of the results, children's fear of being tested, the risk of blood poisoning, and the time taken to wait for the tests to be taken. Therefore, research in the field of development of non-invasive blood hemoglobin detection systems and automation of data recording is relevant.

II. UNDERSTANDING OF PROBLEM

The adult body contains 120-160 grams per liter of hemoglobin in the norm. On average, 4-5 grams of iron is in

the body in various forms, of which about 70% is in the composition of hemoglobin, about 5-10% is in the composition of myoglobin, about 20-25% is in the form of reserve iron and no more than 0.1% is in the blood plasma [10]. Correspondingly, hemoglobin is in the range of 2.8-3.5 grams.

Methods of determining the level of hemoglobin by the amount of iron are based on the fact that in all four chemical groups of hemoglobin molecules, the amount of iron is 0.347%, therefore, the amount of iron can be determined by the level of hemoglobin in the blood.

The determination of the level of iron in the blood, the normal value of which in the human body is in the range of 10 to 30 micromoles per liter, should be determined with some shift, namely, taking into account the pathology. Thus, the operating range of the device should be at least 1 to 50 micromoles per liter.

So, the amount of the iron in the blood is very small from the point of view physical measurements; and it must be detected.

III. PROPOSED TECHNOLOGY

It is proposed to use the ability of iron as a ferromagnetic to be attracted by a magnetic field. The system consists of external measuring unit, which does not assume the invasiveness of the human tissue under the test. The human tissue is affected by a pulsed magnetic field and simultaneously it is illuminated by external light source. The optical receiver is located on the other side of the human tissue and detects the change in light intensity passing through the tissue.

A simplified scheme of the device operation is shown in Fig. 2 and Fig. 3.

Although this technology is not used in clinical laboratories, it can be used to calculate hemoglobin by using the data from the proposed device.

Since the amount of iron in a healthy person's body and a person with anemia is not significantly different from the quantity of iron in the body, there are system's disadvantages associated with this. First, it is the presence of external "noise," caused by external illumination, which will make the error in the readings of the system. Second, rather small values of iron will correspond to a weak amount of changing of light force passing through the tissue.

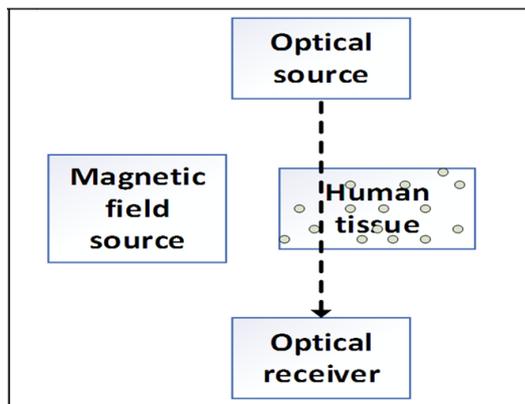


Fig. 2. Principle of device operation in case of external magnetic field presence.

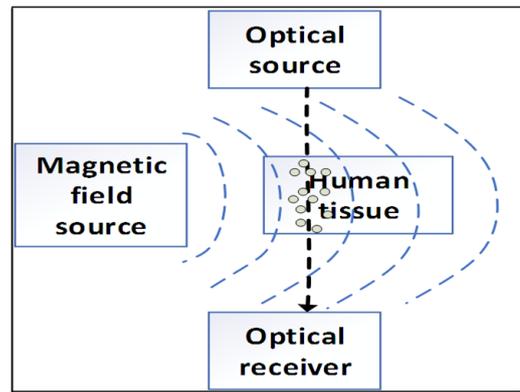


Fig. 3. Principle of device operation in case of external magnetic field absence.

These disadvantages will be taken into account in the design process of the device develop in such a way as to minimize these errors or take them into account under any external environment situation.

Initially, two questions were identified as problems. First, it is the problem of magnetization of iron in the human body and, second, it is the problem of sensitivity of the optical radiation receiver.

The first problem was solved by analyzing reliable fundamental literature, the results of which revealed that iron ions Fe^{2+} and Fe^{3+} [11, 12], whose compounds are subject to magnetic field influence, are present in the organism to a large extent.

For useful signal catching the principle of synchronous detection is used, and it is necessary to determine the required value of the light flux for the source of optical radiation. It is also necessary to determine the strength of the magnetic field created by the electromagnet, as well as to select the frequency of exposure to electromagnetic fields on the body.

Actual values of iron level in blood will be determined on the basis of full-scale experiments, and the obtained values of voltages will be assigned their own values of iron level in blood.

IV. DEVICE DESCRIPTION

In order to the capturing of useful signal, the principle of synchronous detection is used, and it is necessary for determining the required value of the light flux for the source of optical radiation. Also, it is necessary for determining the strength of the magnetic field created by the electromagnet, and for selecting the frequency of exposure of the electromagnetic field on the body.

For creation of the interface of the user convenient for the person, the software including a database of indications will be created.

A detailed scheme of the device, which operates according to the proposed method, is shown in Fig. 4 [13].

The device should consist of two blocks, the primary transducer proper, which converts the changing transparency of human body tissue under the influence of an external applied magnetic field into an electrical DC signal.

The second unit is a computing device with integrated power supply unit.

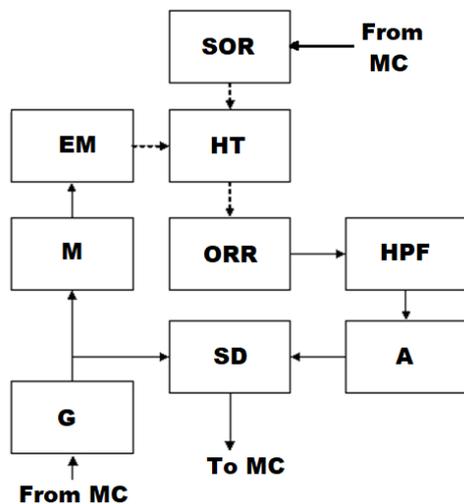


Fig. 4. Detailed structural scheme of the device.

Since it is assumed that the value of electric power applied to the electromagnet should be measured in units and even tens of watts, the power supply of the device from the autonomous current (battery) is inexpedient. For this reason, the device will be powered by AC 220 V 50 Hz.

The primary transducer will contain an electromagnet and an optocoupler placed on an ergonomic device wrapped around the thumb and index finger of the human hand, so that the illuminated area of human tissue falls on the membrane between the thumb and index finger.

The generator (G) forms a pulse that is fed to the synchronous detector (SD) and simultaneously to the modulator (M).

The modulator generates a current flowing through the electromagnet (EM), which creates a magnetic field that affects human tissue, while the optical source of radiation (SOR) forms the light force continuously.

The light passing through human tissue (HT) is modulated in amplitude by the accumulation of red blood cells containing hemoglobin. However, these changes in the luminous flux on the optical radiation receiver (ORR) are extremely small. This level of modulation is in several orders of magnitude less than the average illumination of the optical radiation receiver, created by the constant light passing through human tissue, and external illumination (parasitic illumination).

The first way of eliminating this imbalance is the use of the high-pass filter (HPF), which transmits an alternating signal due to the modulation of the illumination due to the pulsed magnetic field. The spectral components (close to the direct current) caused by the average illumination in the high-pass filter are suppressed.

However, in the signal at the output of the high-pass filter there can be also spectral components lying near the frequency of magnetic field influence. It is necessary to understand this frequency will not be high, may be several hertz. Such value of frequency is caused the fact that the time constant of filling the tissue with blood is quite large. Therefore, any movement of human tissue will result in the appearance of components in the spectrum of the signal that are commensurate with the frequency of exposure to the magnetic field.

Low frequency amplifier (A) forms on its output gained sum of useful and parasitic components. Consequently, the amplified spectral components have not to be detected directly, because parasitic components in amplitude will be many times larger than useful one.

It is proposed to use the synchronous detector. Due to the presence of an integrating chain the synchronous detector can distinguish between useful and parasitic signals with a signal-to-noise ratio of up to -100 dB. For example, the AD630 synchronous detector from Analog Devices can be used [14].

In this way, a direct current signal is obtained at the output of the synchronous detector. Signal detection is clocked by the reference oscillator signal, and it is proportional to the changing of light intensity only at the input of receiver of optical radiation. These changes are caused by the affects of an external magnetic field on human tissue.

However, it is necessary to understand which values of light intensity and magnetic field should be used. These values will be determined experimentally.

The device operation is managed by the control unit, the main part of which is microcontroller (MC), as it is shown in Fig. 5.

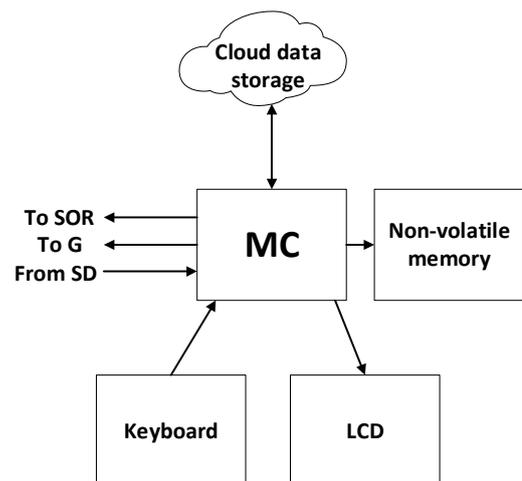


Fig. 5. Structure of control unit.

The primary transducer described above is connected to a control unit with a three-signal-wire remote cable. One wire must be placed in the shield. Besides the signal wires the power wires have to be included into the cable as well. The cable has to be connected to control unit via small connector with seven pins.

The user interface must have an LCD display showing the iron content in micromole per liter and can also contain addressable statistics (for each client).

The device will be controlled by means of the keyboard. Availability of non-volatile memory (NVM) will allow entering statistics of diseases and store data. The data can be subsequently sent to the server of the medical facility. This will eliminate the need for the patient to be present in person at the doctor's office if there is no need to be present. At this stage of the research, the device will not be communicated with a personal computer.

V. CONCLUSION

The problem of controlling of the level of hemoglobin in blood was considered in the paper. Existing methods of determining the level of hemoglobin were discussed as well. The main drawbacks of the existing systems were determined. These methods were based on their invasiveness, difficulty in getting the results, as well as the duration of data processing.

The new method of determining the level of iron in the blood was proposed. The method is based on the properties of the iron, as ferromagnetic, to be attracted by external magnetic field. At the same time, translucent working area of the human tissue surface will change the optical transparency, which is proportional to the content of iron in a blood.

Possible risks of system error and efforts for its reducing were described. The structural scheme of the device for non-invasive determination of iron level in blood was proposed. The device consists of two units: primary transducer and control unit.

Special arrangements in primary transducer design let decreasing the negative influence of parasitic illumination of optical receiver. Synchronous detection of the sum of useful and parasitic signal ensures the selecting of useful signal with the signal/noise ratio up to -100 dB.

The control unit implements the obtaining measuring data on hemoglobin-in-blood presence very fast, within few seconds. Device ensures fast and non-invasive determination of hemoglobin. The risk of the blood infection is null, at least no more than causal use of any other thing by man.

Device has possibility of using data to be transferred to medical institutions for statistical result obtaining without personal presence.

In the perspective it is supposed to research the dependence of parameters determined by our device on other parameters of the body. For example, in articles [15, 16] there are described the methods of monitoring the condition of patients with diabetes by determining the level of hemoglobin. Such researches give us possibility to presume that magnetic field may change some body parameters which influenced on diabetic level. It is planned to study the

influence of magnetic field on the parameters of the organism.

REFERENCES

- [1] Big Medical Encyclopedia, Ed. B. V. Petrovsky, 3rd Ed., Vol. 1-30, Moscow, "Soviet Encyclopedia," 1974, Vol. 1, P. 518-523.
- [2] H. M. Ranney, V. Sharma, "Structure and Function of Hemoglobin," In Beutler E Lichtman M et al (eds). Willimas's Hematology. 6th edition. McGraw Hill, 2000: 345-53.
- [3] Big Medical Encyclopedia, Ed. B. V. Petrovsky, 3rd Ed., Vol. 1-30, Moscow, "Soviet Encyclopedia," 1974, Vol. 5, P. 154-156.
- [4] N. Harris, G. Devoto, J. Pappas, "Performance evaluation of the ADVIA 2120 hematology analyzer: an international multicenter clinical trial," *Lab Hematol* 2005; 11: 62-70
- [5] J. Rosenbilt, C. Abreu, L. Sztlering, "Evaluation of three methods for hemoglobin measurement in a blood donor setting," *Sao Paulo Med J* 1999; 117: 108-12
- [6] S. Sharp, S. M. Lewis, S. K. Williams, "Evaluation of the HemoCue B-haemoglobin photometer," Medical Devices Agency evaluation report MDA/95/21 1995. London HMSO
- [7] A. Medina, C. Mundy, J. Kandulu, L. Chisuwo, "Bates Evaluation and costs of different haemoglobin methods for use in district hospitals in Malawi," *J Clin Pathol* 2005; 58: 56-60
- [8] J. Ray, J. Post, C. Hamielec, "Use of rapid arterial blood gas analyzer to estimate hemoglobin concentration among critically ill adults," *Critical Care* 2001; 672-75
- [9] T. V. Rusova, G. A. Ratmanova, O. V. Kozlov, "Diagnosis of iron deficiency anemia in children," *Zemsky Doctor*, 2011, No 5 (9), P. 13-16.
- [10] P. Lenkin, A. A. Smetkin, A. Hussein, "Continuous monitoring of hemoglobin by pulse oximetry after cardiac surgery," *Russian Journal of Anaesthesiology and Reanimatology*, 2016, No 61 (5), P. 329-333.
- [11] Big Medical Encyclopedia, Ed. A.N. Bakulev., Ed. 2nd Ed., Vol. 1-33, Moscow, "Soviet Encyclopedia," 1958, Vol. 6, P. 717-730.
- [12] Big Medical Encyclopedia, Ed. B. V. Petrovsky, 3rd Ed., Vol. 1-30, Moscow, "Soviet Encyclopedia," 1974, Vol. 8, P. 54-59.
- [13] A. A. Azarov, E. I. Shirokova, "The method of determining the level of iron in a blood," Patent application # 2018139008, Russian Federation, IPC A61B 1/06 from 08.11.2018.
- [14] AD630 [Electronic resource], Access mode: <http://www.alldatasheet.com/datasheet-pdf/pdf/48099/AD/AD630.html>.
- [15] A. V. Ilin, M. I. Arbuzova, A. P. Knyazeva, "Glycated hemoglobin as a key parameter in monitoring diabetes patients. Optimal research organization," *Diagnostics control and treatment*, 2008, P. 60-64.
- [16] Demir Aslıhan Dilara, Durmaz Zeynepüllya, Kılınc Çetin, Güçkan Rıdvan1, "Correlation Between Red Blood Cell Distribution Width and Glycated Hemoglobin in Diabetic and Nondiabetic Patients," *Russian Open Medical Journal*, 2016, P. 301-304.

The Using of Electronic Document Management Tools of Technical Documentation for the Assessment of the Life of the Train Traffic Control Devices

Dmitry V. Sedykh,
Chief engineer of Group of Companies «*IMSAT*»,
engineer at “Automation and Remote Control on Railways”
Department,
Emperor Alexander I
St. Petersburg State Transport University
Saint-Petersburg, Russia
sedyhdmitriy@gmail.com

Michael N. Vasilenko
PhD, professor, Professor at
“Automation and Remote Control on Railway” Department
Emperor Alexander I St. Petersburg State Transport University
Saint-Petersburg, Russia
vasilenko.m.n@gmail.com

Andrei Belyi,
PhD, associate professor, cathedra “Bridges” chief,
Emperor Alexander I St. Petersburg State Transport University
Saint-Petersburg, Russia
andbeliy@mail.ru

Denis V. Zuyev,
PhD, Director general of Group of Companies
«*IMSAT*», Head of CAD Center at “Automation
and Remote Control on Railways” Department,
Emperor Alexander I
St. Petersburg State Transport University
Saint-Petersburg, Russia
zuevdv@gmail.com

Michael A. Gordon,
Chief specialist of the Institute “*Giprotranssignalsvyaz*” –
Department of JSC “*Roszheldorproject*”
Saint-Petersburg, Russia
gordon_ma@mail.ru

Abstract — A number of railways have posed repeatedly the requirements to provide a capability of shifting towards repair and recovery technology for maintenance of the systems due to trouble-shooting and forecasting of the device statuses and recording of the train traffic control devices based on their actual operation hours. At present, the integrated monitoring hardware permit to shift towards the above mode only for a number of devices equipped with the operation monitoring facilities.

The goal of this work is developing a methods for calculating the resource based on existing technical documentation and data monitoring systems, without the additional cost of entering the source data. To this end, a method has been developed for constructing a model of the operation of a technical system for circuit solutions and protocols for its operation. Scientific innovation consists in developing a universal method that does not depend on the type of technical system, but is entirely determined by its description in a universal format.

The paper proposes a method based on the graph and binary logic theory to receive data on the operation hours of all the system devices based only on available information and technical documentation for the system only. The proposed solution permits to obtain real data on the operation hours of automation devices without additional hardware costs, requiring no fitting of every system component with monitoring equipment. The integrated principles of technical documentation generation in a special format provide the prerequisites for intelligent processing of the data. After a traffic control system design was made in a special technical documentation format, a method has been developed to automate building of a real system model and analyze its operation. Owing to a combination of a computer modeling system and a record of real process operations implemented on a specific train traffic control system, a method has been

developed to obtain data on operation of all the components at any point in time. And when an expected train schedule is built into the developed model, it is possible to plan a maintenance schedule in advance for a forthcoming period based on the actual condition without integration of new special hardware. **Keywords**— special format, train traffic control, logic diagrams, online documentation, industry format of technical documentation, modelling, logic modelling.

I. INTRODUCTION

At present, there is no troubleshooting and monitoring system capable of picking up the data from all railroad automation and telemechanic (RAT) devices [1-2] that gives no chance to reject totally regular replacement of the automation and telemechanic components under a schedule without regard to their actual operation hours. All diagnostic monitoring systems [3-6] suppose that data pickup devices should be installed on a monitoring component. It is impossible and unreasonable in technical terms to install a monitoring instrument on every separate control system component, though correct data on the operation hours of every component is necessary to forecast its life. It means that engineering solutions are required, which will permit to assess the operation hours of RAT devices without installation of an additional troubleshooting package.

The central purposes set to solve the problem are to:

- shift towards maintenance and inspection of the traffic control devices on condition rather than under a schedule;
- identify the actual life of traffic control devices without additional costs;
- forecast the life of traffic control devices in the future.

Autors proposes to use modelling of the system operations based on technical documentation. Modelling should identify the life of every component. Input actions on the developed model are a traffic schedule and a record of process operations on the system. This could be an expected traffic schedule for provisional maintenance planning or correct information from the microprocessor-based traffic control system or operation monitoring system records.

II. SPECIFIC FEATURES OF TECHNICAL DOCUMENTATION PRESENTATION

So far, a suite for dealing with technical documentation has been developed for the Russian railways [5]. The software performs an integrated set of tasks, starting from an electronic archive and electronic document flow system and ending with an industry CAD system concerning railway automation with functions of an expert system. The system stores all the technical documentation for RAT devices, which could be used as a data source.

All the required technical documentation is presented in an Russian industry format of technical documentation (IF-TD) [8, 9]. This format like as other industrial formats in the world (RailML® and other) [10-15]. The IF-TD (Fig. 1, 2) contains not only the drawing information (drawing by SVG format)[16], but also a model of the depicted device or a separate component by means of Extensible Markup Language (XML) [17]. Every component of the drawing is described by a full set of parameters, which permit to identify uniquely a specific set on the real system as well as obtain its technical specifications from a file.

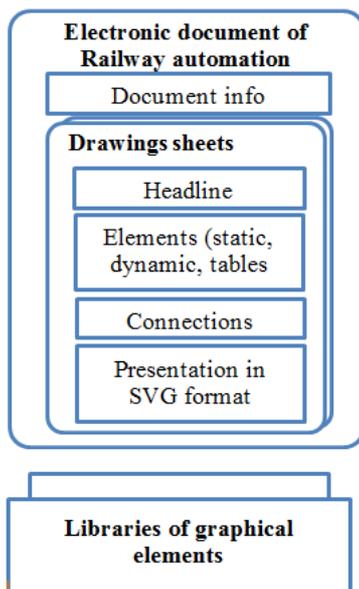


Fig. 1. IF-TD structure.

Any file (Fig. 3) in IF-TD contains a description:

- document info;
- drawings sheets with Headline;
- elements (static, dynamic, tables etc.) ;
- connections between the elements;
- presentation in SVG format.

File in IF-TD provides substantially a model of the depicted device or part of the system.

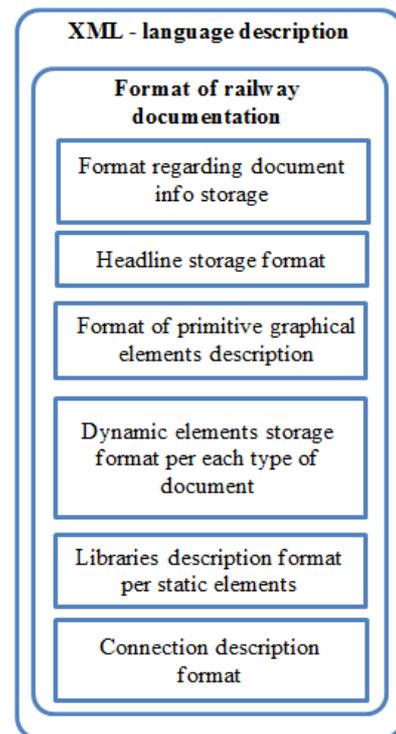


Fig. 2. IF-TD structure.

```
<Document Version = "1.00.83" LastID = "3" LastPageID = "2"
xmlns:bsl="http://www.imsat.spb.ru/libs/ashapes.IMSATAutoShap
esLibrary" xmlns:e="http://www.imsat.spb.ru/editor/v1.0"
xmlns:cp="http://www.imsat.spb.ru/editor/v1.0/custprop"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.imsat.spb.ru/editor/v1.0">
  <Page PageID = "1" Format = "A4" Width = "210" Height =
  "297">
    <bsl:rect id = "1" class = "style5 style2" x = "9.0791334" y =
    "21.608729" width = "12.964583" height = "12.170833" />
    <bsl:path id = "2" class = "style5 style2" transform = "matrix (1 0
    0 1 36.68998 41.54666)" d = "M0 0 C-0 -3.0868056 -1.2347223 -
    14.772571 0 -18.520834 S4.4538193 -22.489584 7.4083333 -
    22.489584 S16.095486 -22.974653 17.727083 -18.520834
    S14.022916 -1.1906251 17.197916 4.2333331 S39.907986
    9.8777781 36.777084 14.022917 S5.8649302 26.678818 -1.5875
    29.104166 S-5.291667 32.146873 -7.9375 28.575001 S-15.279688
    13.714235 -17.4625 7.6729169 S-21.607639 -6.3940978 -
    21.128555 -7.7670972" />
    <bsl:path id = "3" class = "style5 style2" transform = "matrix (1 0
    0 1 22.846445 23.507903)" d = "M-7.2850204 -1.8991735
    L4.8477008 -20.59041 L13.843534 18.038757" />
    <Links>
      <Link t1 = "id (1) / Bottom" t2 = "id (2) / P2" />
      <Link t1 = "id (1) / Top" t2 = "id (3) / P1" />
      <Link t1 = "id (1) / Geometry" t2 = "id (2) / P2" />
      <Link t1 = "id (2) / P1" t2 = "id (3) / P2" />
    </Links>
  </Page>
```

Fig. 3. Description of element connections

III. CONSTRUCTION OF UNIFIED MODEL BASED ON TECHNICAL DOCUMENTATION

As mentioned above, technical documentation contains all the information required for construction of a unified system model. All that is left to do is to remove the data that is not required for the problem and combine a lot of sheets (drawings) into a unified model.

Construction of a model could be split provisionally into three stages:

- establish a working space from the diagram sheets;
- construct electrical units of the diagram, which govern electrical connections among the elements.
- identify the devices used in the diagram from separated parts on the drawing sheets.

A working space designed for data control creates and stores the data from the diagram sheets describing the system operation principles. Electrical units are presented both on separate sheets as well as can continue on other design sheets, so a process of obtaining and combining the units takes place throughout construction of a model.

Instruments can be presented on the diagrams as a lot of parts and outputs (for example, a relay winding and contact) depicted often on different sheets of a diagram. To identify correctly the instruments, it is necessary to identify all the parts of devices from all the sheets of a diagram.

As a result, it is possible to state that a model constructed automatically from the technical documentation contains two main components that are necessary to solve the posed problem:

- set of electrical units of the system;
- set of devices with a set of parameters.

The units and instruments are interconnected with the instrument outputs, which are contained in a description of a specific instrument and operate as connection points with a specific electrical unit. A set of devices include both the controlled devices as well as power supply and input signal sources to set the model operation parameters.

A produced unified model of the system permits also to solve a lot of other problems. At present, the model in question is applied to problems of controlling the circuit designs and technical documentation:

- check of device block diagram plotting;
- check of device assembly diagram plotting;
- control of matching between different diagram types;
- check of technical documentation for compliance with rated values;
- check of technical documentation after operation of certain modules for generation of the same;
- control diagram after recognition[18-19].

IV. AUTOMATION SYSTEM OPERATION MODELLING METHODS

A produced system model based on the technical documentation provides a basis for construction of a logic operation model of train traffic control systems. Specification of the input actions as status setting of certain model components makes it possible to have an operation modelling

process of all the system in block. It is necessary to consider that modelling can be carried out as two types: modelling of electrical processes of all the system's operation, including transient phenomena, and logic modelling. Modelling of electrical processes of the system (analog-digital modelling). This permits to carry out: a full functional check of the system, diagnostics of the operation modes, and many other tasks, but is quite labor intensive.

For the purposes of recording the operation hours of devices, logic modelling tools are sufficient. In case of logic modelling, a status of all the system components is presented as two potential statuses of «1» and «0», i.e. «on» and «off». At the same time, all the system is presented as a set of Boolean equations. Every equation of the system describes a way to turn on the element at a given time step. A solution to all the system of equations governs the system status at a given point in time. A set of Boolean equations is set up automatically based on a plotted graph, i.e. the enabling circuits of every system element are identified at every point in time. A presented system of equations describes all potential operation of the system to identify the status at every point in time.

Methods and tools to convert a file describing a system drawing in a special open data format into a computer model that is suitable for supply of actions on the same concerning a train traffic schedule or based on the actual traffic have been developed in this paper. A developed methodology integrating the data of monitoring systems permits to transmit this automatically to the obtained model in order to synchronize the model performance with the operating devices.

Modern computing facilities permit to solve the system of equations almost on any timescale from real time to computations for future years. A modelling outcome depends on the specified timescale to perform one time step of a real time system.

The work outcome may be output in any form that is convenient for the user: operational tables of all the system components, a timing chart of the device behaviors (Fig. 4), and many other.

The model as such can operate in different modes depending on an input data setting method:

1. Manual mode, where the actions are specified manually by the user from a control unit analog.
2. Traffic schedule setting. A traffic schedule and a set of standard actions implementing the schedule are supplied to the model, on which basis the model performs. The mode can apply to develop a provisional maintenance schedule for the devices and to develop plans for exchange of the devices from the installation points; from a point featuring a low number of actuations to that with a high number of actuations.
3. Mode of integration with supervisory control systems[20-23]. In that mode, all the input actions are supplied from the troubleshooting and monitoring systems [24-26]. The model operates in parallel with the real system to count the life of all the instruments in real time.

The manual mode can be applicable to develop automatically simulators of real train traffic control systems based only on their technical description as well as to monitor correctness of circuit designs for automatic examination of the circuit designs of train control systems.

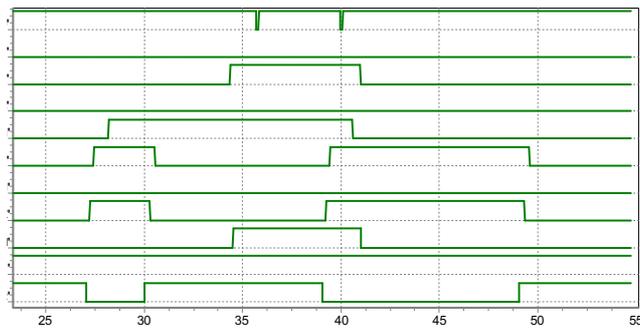


Fig. 4. Timing charts of device (instrument) behaviors

V. CONCLUSION

This paper develops tools for use of the electronic technical documentation to find out the actual operation hours of devices without installation of dedicated monitoring facilities on every element of the system. It is shown that this do not require development of separate dedicated models, but it is sufficient to take only the technical documentation describing the system. The work proposes an engineering solution for automatic modelling of the operating systems and provides an overview of the methods to produce a unified model from the technical documentation that can bring a high cost advantage for a shift towards maintenance based on the actual condition of the systems consisting of many components, which are maintained currently based on the operation life only. The method permits to obtain automatically an operating system model based just on its technical documentation and start to count the life of elements based on the available data, without expenses for integration of the dedicated engineering systems of component troubleshooting.

In this paper, methods for calculating the resource are developed on the basis of available technical documentation and data from monitoring systems, without the additional cost of inputting initial data and installing additional systems.

The developed methods make it possible to construct a model of the operation of a technical system of any type only according to technical documentation.

REFERENCES

- [1] G. Theeg, S. Vlasenko (eds.) "Railway Signalling & Interlocking: International Compendium", 2nd revised edition, PMC Media House GmbH, 2017, 458 p.
- [2] P.D.F. Conradiel, C.J. Fourie, P.J. Vlok, and N.F. Treurnicht, "Quantifying System Reliability in Rail Transportation in an Ageing Fleet Environment," *South African Journal of Industrial Engineering*, August 2015, vol. 26(2), pp. 128-142.
- [3] V.P. Molodtsov, and A.A. Ivanov, "Dispatch control and monitoring system for railway automation and remote control devices" (in Russ.), St. Petersburg: Petersburg State Transport University, 2010, 140 p.
- [4] D.V. Efanov, "Concurrent checking and monitoring of railway automation and remote control devices" (in Russ.), St. Petersburg: Emperor Alexander I St. Petersburg State Transport University, 2016, 171 p.
- [5] A.A. Ivanov, A.K. Legon'kov, and V.P. Molodtsov, "Data Transmission from APK-DK Device of Rail Crossing Under the Absence of Physical Link and Clock Duty" (in Russ.), *Automation on Transport*, 2016, vol. 2, issue 1, pp. 65-80.
- [6] D.V. Efanov, V.V. Khóroshev, G.V. Osadchy, and A. Belyi, "Optimization of Conditional Diagnostics Algorithms for Railway Electric Switch Mechanism Using the Theory of Questionnaires with Failure Statistics," *Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2018)*, Kazan, Russia, September 14-17, 2018, pp. 237-245. DOI: 10.1109/EWDTS.2018.8524620.
- [7] K. Winter, W. Johnston, P. Robinson, P. Strooper, and L. Vanden Berg "Tool Support for Checking Railway Interlocking Designs," *Proceeding of the 10th Australian Workshop on Safety Related*

- Programable Systems (SCS'05), Australian Computer Science Communications, 2005, pp. 101-107.
- [8] D.V. Sedyh, D.V. Zuev, and M.A. Gordon, "Industry Framework for Technical Documentation for Railway Automation and Remote Control Devices. Part 1: Concept of Design," *Automation on Transport*, 2017, vol. 3, issue 1, pp. 112-128.
- [9] D. Sedykh, M. Gordon, and D. Efanov, "Computer-Aided Design of Railway Signalling Systems in Russian Federation," *Proceedings 2018 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, Moscow, Russia, May 15-18, 2018, pp. 1-7. DOI: 10.1109/ICIEAM.2018.8728630
- [10] RailML – a standard interface for railway data [Online]. Available: <https://www.railml.org/en/>.
- [11] A. Hlubuček, "RailTopoModel and RailML 3 in overall context," *Acta Polytechnica CTU Proceedings*, 2017, vol. 11, pp 16-21.
- [12] A. Nash, et al. "RailML - a standard data interface for railroad applications," *Proceedings of the Ninth International Conference on Computer in Railways (Comrail IX)*, Dresden, Germany, 2004.
- [13] Z. Łukasik, W. Nowakowski, and T. Ciszewski, "Definition of data exchange standard for railway applications," *Prace Naukowe Politechniki Warszawskiej - Transport*, 2016, no. 113, pp. 319- 326.
- [14] M. Bosschaart , et al. "Efficient formalization of railway interlocking data in RailML," *Information Systems*, April 2015, vol. 49, pp. 126-141.
- [15] T. Ciszewski, W. Nowakowski, and M. Chrzan, "RailTopoModel and RailML – data exchange standards in railway sector," *Archives of Transport System Telematics*, November 2017, vol. 10, issue 4, pp. 10-15
- [16] SVG – Scalable Vector Graphics [Online]. Available: <https://www.w3.org/Graphics/SVG/>.
- [17] XML – Extensible markup language [Online]. Available: <https://www.w3.org/XML/>.
- [18] E.A. Blagoveschenskaya, D.V. Zuev, V.V. Garbaruk, V.A. Gerasimenko, D.V. Sedykh, and D.S. Kunets, "Application of Convolutional Neural Networks for Pattern Recognition Circuits of Railway Automatics. Specifics of This Application," *Proceedings of 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM)*, St. Petersburg, Russia, May 24-26, 2017, pp. 434-435.
- [19] N. Hage and B. Wagner, "A new function block modeling language based on Petri nets for automatic code generation," *IEEE Trans on Industrial Informatics*, vol. 1, pp. 226-237, 2005.
- [20] V.V. Khóroshev, D.V. Efanov, and G.V. Osadchii, "Ways of Development of Periodical and Continuous Monitoring Means for Automatic Devices on Marshaling Yards," *Proceedings of 1th International Russian Automation Conference (RusAutoCon)*, Sochi, Russia, September 9-16, 2018, pp. 1-5, doi: 10.1109/RUSAUTOCON.2018.8501720.
- [21] D. Efanov, G. Osadchy, and D. Plotnikov, "Average Number of Orders Calculation Concerning Diagnostic Test of Measuring Controllers During Permanent Monitoring Performance Based on Stationary Model of Queueing System," *Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS'2018)*, Kazan, Russia, September 14-17, 2018, pp. 660-670, doi: 10.1109/EWDTS.2018.8524638.
- [22] D. Efanov, G. Osadchy, D. Sedykh, D. Pristensky, and D. Barch, "Monitoring System of Vibration Impacts on the Structure of Overhead Catenary of High-Speed Railway Lines," *Proceedings of 14th IEEE East-West Design & Test Symposium (EWDTS'2016)*, Yerevan, Armenia, October 14-17, 2016, pp. 201-208.
- [23] D. Efanov, D. Pristensky, G. Osadchy, I. Razvitnov, D. Sedykh, and P. Skurlov, "New Technology in Sphere of Diagnostic Information Transfer within Monitoring System of Transportation and Industry," *Proceedings of 14th IEEE East-West Design & Test Symposium (EWDTS 2017)*, Novi Sad, Serbia, September 29 – October 02, 2017, pp. 231-236.
- [24] S. Mokrousov, N. Naish, S. Demjanenko, V. Bykadorov, and S. Ramazanov, "Analysis and evaluation of the life cycle cost of technical rail systems (for example, a wheel pair)," *TEKA Commission of Motorization and Power Industry in Agriculture – 2013*, Vol. 13, No 3, pp. 152-161.
- [25] V. Vyatkin and H.-M. Hanisch, "Verification of Distributed Control Systems in Intelligent Manufacturing," *Journal of Intelligent Manufacturing*, vol. 14, pp. 123-136, 2003.
- [26] D. Efanov, G. Osadchy, and D. Sedykh, "Protocol of Diagnostic Information Transmission via Radio Channel Concerning Health Monitoring of Infrastructure of Russian Rail Roads," *Proceedings of 3ed International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, St. Petersburg, Russia, May 16-19, 2017, pp. 1-6.

Automation of layout design of spiral conical scans

Marina V. Byrdina
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
byrdinamarina@mail.ru

Lema A. Bekmurzaev
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
Bekmurzaev.l@yandex.ru

Mikhail F. Mitsik
Mathematics and applied informatics
Don State Technical University
Rostov-on-Don, Russia
m_mits@mail.ru

Dmitry B. Kelekhsaev
International logistics systems and complexes
Platov South-Russian State
Polytechnic University (NPI)
Novocherkassk, Russia
d-kelekhsaev@mail.ru

Anatoly I. Kondratenko
Engineering design
Russian State Agrarian University –
Moscow Timiryazev Agricultural Academy
Moscow, Russia
ai_kondratenko@mail.ru

Abstract—In this paper the authors solve the problem of dense layout on the infinite plane of the spiral scans of the side surface of a straight circular truncated cone. The boundaries of the spiral scans are arcs of circles and spirals of Archimedes. A full scan of the side surface of the cone is a single part. The approach of finding a dense layout by means of implementation of transformations of rotation of figures and their parallel transfer in the Maple 2015 is offered. As an example of the surface, the surface of a conical spiral antenna or the surface of a woman's skirt can be considered. It is shown that for the proposed layout of the spiral scan the proportion of cuttings is equal to 5.3%, and for straight conical scans of the same lateral surface is 10.2%. Thus, the proposed method of layout of the spiral scans in the plane allows to reduce the proportion of cuttings by almost 5%

Keywords— dense layout, spiral scans, infinite plane, straight cone, model of a female figure and antenna, the percentage of cuttings

I. INTRODUCTION

Various products and devices based on conical surfaces with planar and space involute curves have been widely used in science and technology. The analysis of works in this direction showed that such approaches are used in modelling and simulation of conical antennas with two helical copolarization layers for the transmission of high-frequency signals [1, 2], when creating a tooth surface treatment technique for a conical gear with a tooth profile curve in form of the spiral curve [3], in the development of engineering methods for designing flat and spatial logarithmic spirals for modeling conical gears, conical spiral gears, spiral springs [4], etc.

Due to the property of flattening the cone surface has wide applications in engineering and technology. Depending on the application and purpose, conical surfaces can be made of various fabrics, carbon fiber, metals, plastics, etc.

Analysis of the results of the introduction of computer-aided design systems in production shows that the automation of cutting of product surfaces not only reduces costs and improves their quality, but also significantly increases the percentage of material usage, rational consumption of which is particularly relevant for enterprises using expensive raw materials [5, 6, 7].

The goal of this work is to develop a method of the spiral scans dense layout on an infinite plane for the lateral surface of a truncated straight circular cone with a cuttings

fraction less than that of traditional direct scans dense layout methods. Key objectives: 1) calculate the percentage of cuttings for the direct scans layout; 2) determine the percentage of cuttings for the spiral scans dense layout and show the advantages of such a layout; 3) construct spiral scans on a plane, as well as dense layouts of spiral scans and direct scans, using Maple programming environment.

For the traditional design of a cone's lateral surface scans on an infinite plane, the scheme illustrated in Figure 2 is used. The novelty of the suggested method lies in the development of a new form of a cone's lateral surface - spiral scans - which allow to reduce a cuttings fraction from 10.2% to 5.3%. Modeling of a cone's lateral surface in the form of spiral scans also allows to find new design solutions when designing clothes and conical spiral antennas.

II. PERCENTAGE OF CUTTINGS FOR STRAIGHT SCAN

Cone surfaces are widely represented in various industries: light and manufacturing industries, sphere of services, engineering, etc.

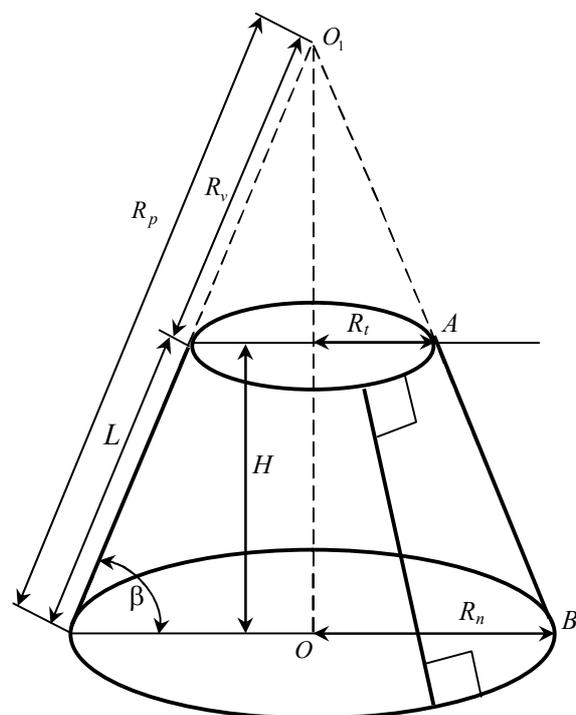


Fig. 1. Modeling of a straight circular truncated cone

Conical surfaces have the property that they all unfold on a plane without folds and breaks, respectively, from the part on a flat sheet of material, deforming it, you can get a conical surface shape (Fig. 1).

The problems of optimal cutting were solved by a number of Russian and foreign scientists [8, 9, 10]. Many of works were associated with the regular layout, i.e. the positioning of the same figures in the endless strips [11, 12].

As an example of designing a conical scan, the parameters of scan of the side surface of a straight circular truncated cone will be found for a skirt on a female figure of a typical physique, the initial data are given in Table 1.

TABLE I. ANTHROPOMETRIC DATA OF A TYPICAL FEMALE FIGURE

Measures, cm		
1	Heigh	164
2	Waist circumference C_w	76
3	Projection distance from waist to thigh h	20
4	Bust circumference C_b	96
5	Hip circumference C_h	104
6	Length of skirt L	53

Similarly, the surface of the conical antenna can be designed, for example, with the same parameters of the cone.

We define the parameters of the straight cone and the corresponding classical scan by formulas

$$R_t = \frac{C_w}{2\pi}, R_b = \frac{C_h}{2\pi}, \beta = \arctg\left(\frac{h}{R_b - R_t}\right), \quad (1)$$

$$\alpha = 2\pi \cos \beta, R_v = \frac{R_t}{\cos \beta}. \quad (2)$$

$$R_p = R_v + L; R_n = R_p \cos \beta, \quad (3)$$

$$H = L \sin \beta \quad (4)$$

The values of the parameters of the classical straight circular truncated cone scan are given in table 2.

TABLE II. VALUES OF STRAIGHT SCAN PARAMETERS

Measures, cm		
1	Radius of waist R_t	12.1
2	Radius of hip R_b	16.55
3	Radius of length of skirt R_n	23.62
4	Height of skirt H	51.73
5	Angle between generator and cone base β , rad	1.35
6	Length of generator L	53
7	Length of generator of upper part of cone R_v	55.62
8	Central angle of scan α , rad	1.37

To determine the percentage of material usage with a dense regular layout of straight conical scans, we present the visualization of the layout in the Maple environment (Fig. 3) for the above values of scan parameters.

In Figure 2 it is easy to see that the main rectangle $ABCD$ also contains exactly two copies of straight scans, the area of the main rectangle is equal to

$$S_{rest} = (0,895 + 0,392) \cdot 0,995 = 1,280565 \text{ m}^2 \quad (5)$$

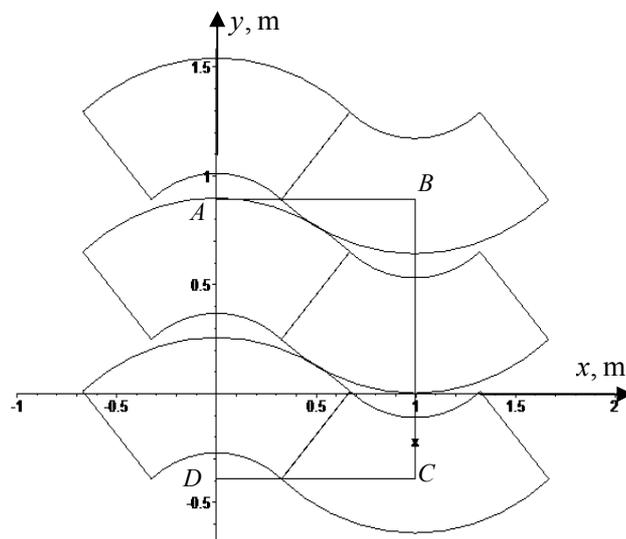


Fig. 2. Dense layout of straight cone scans

The area of one straight scan of the side surface of the cone is determined by the formula

$$S_{str\ scan} = \frac{\alpha}{2} (R_p^2 - R_v^2) = 0,5947 \text{ m}^2. \quad (6)$$

In Figure 2 it is easy to see that in the main rectangle $ABCD$ exactly two scans are placed in the main rectangle $ABCD$, then the effective area of the material usage in the case of a straight conical scan equals to

$$S_{str} = 2 \cdot S_{str\ scan} = 1,18948 \text{ m}^2. \quad (7)$$

The percentage of cuttings is determined by the formula

$$P = \frac{|S_{rest} - S_{str}|}{S_{rest}} \cdot 100\% = 10,24\%. \quad (8)$$

III. PERCENTAGE OF CUTTINGS FOR SPIRAL SCAN

In the work [6] the equations of the boundary of the spiral cone scan in the polar coordinate system were obtained

$$r_{left} = \frac{L}{\alpha \cdot p} \cdot (\varphi - \alpha) + R_v, \alpha \leq \varphi \leq \alpha \cdot (p + 1); \quad (9)$$

$$r_{right} = \frac{L}{\alpha \cdot p} \cdot \varphi + R_v, 0 \leq \varphi \leq \alpha \cdot p;$$

$$r = R_v; r = R_p, \quad (10)$$

where: r – is polar radius, φ – is polar angle;

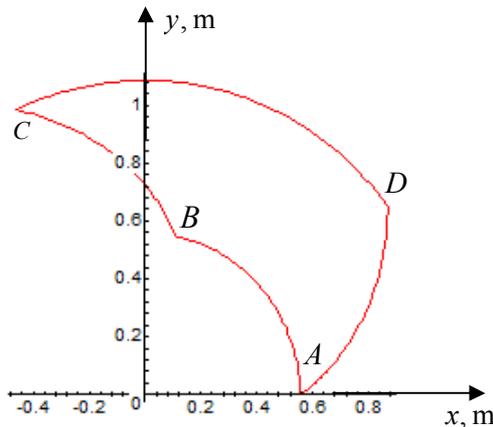
α – central angle of scan;

r_{left}, r_{right} – left and right borders of spiral scan;

p – the number of revolutions of the boundary of the spiral scan on the side surface of the cone.

Let us construct the border of the scans in the form of (3-5) with two values of the parameter p (0.47; -0.47). The visualization of the scans calculated in Maple 2015 for the initial data from Table 1 is shown in Figure 3. Layout, shown in Figure 3 a) we call the left, and in figure 2 b) – the right.

a) $p = 0,47$;



b) $p = -0,47$;

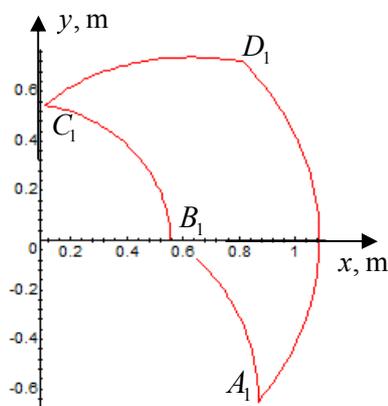


Fig. 3. Dense Scans of cone with data from Table 1

For the horizontal arrangement of the common tangent to the considered scans, we perform the rotation of these scans to the angles β_1 and β_2 , accordingly, as well as the parallel transfer of the scans and their dense layout. Each row consists of left and right scans in turn. For dense scans on an infinite plane, the calculations in Maple 2015 show the following location of the scans and the dimensions of the basic rectangle

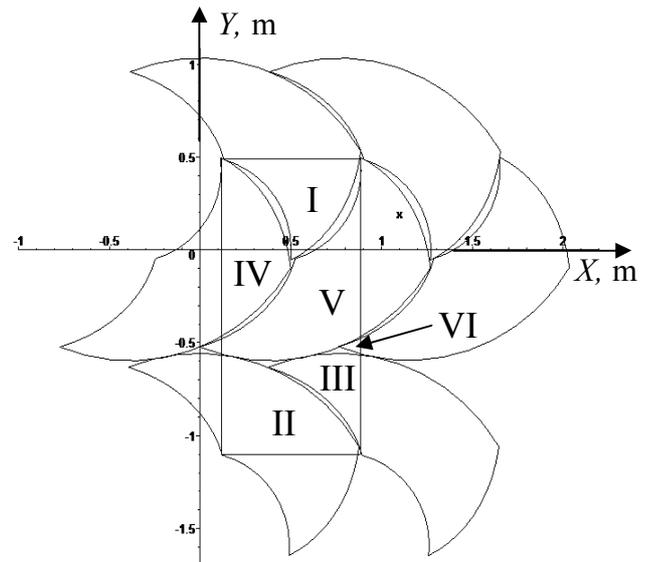


Fig. 4. Example of plotting a dense layout of spiral scans

The rectangle that cuts off the part of the plane that contains all the characteristic elements of the scans is called the basic rectangle. In other words, the main rectangle has the property that joining the same basic rectangles horizontally and vertically will lead to the construction of the entire plane with a dense layout of the scans [13, 14].

Since the second of the left scans is obtained from the first one by a shift of 0.79, the width of the main rectangle is 0.7654 m. Since the left scans of the third row are obtained from the scans of the first row by a shift of 1.59, the area of the basic rectangle is equal to

$$S_{rest1} = 0,79 \cdot 1,59 = 1,217 \text{ m}^2 \quad (11)$$

On the other hand, the basic rectangle includes six parts (along with cabbage). It is easy to see (Figure 4) that combining parts with numbers I, II, III gives the left scan, and combining parts with numbers IV, V, VI gives the right scan. Thus, the useful area of the basic rectangle is the area of two scans.

Let us find the area of each of the scans. All of them are equal-sized figures, since they cover the same straight circular truncated cone. The area of the left scan is calculated in the polar coordinate system by the formula

$$S_1 = \frac{1}{2} \left[\int_0^{\alpha \cdot p} ((a\varphi + R_v)^2 - R_v^2) d\varphi + \int_{\alpha \cdot p}^{\alpha} (R_p^2 - R_v^2) d\varphi + \int_{\alpha}^{\alpha + \alpha \cdot p} (R_p^2 - (a(\varphi - \alpha) + R_v)^2) d\varphi \right] \quad (12)$$

where $a = \frac{L}{\alpha \cdot p}$.

All parameter values are defined in Table 2. We calculate the area of the left scan using the formula (10)

$$S_1 = 0,5947 \text{ m}^2.$$

The basic rectangle includes one copy of the left scan and one copy of the right one, since each of them covers the same side surface of the cone, their areas are equal. Then the area of spiral patterns inside the basic rectangle is equal to

$$S_{spir} = 2S_1 = 1,18945 \text{ m}^2.$$

According to (9), the percentage cuttings for layout of the spiral scans is determined by the formula

$$\Pi = \frac{|S_{rest1} - S_{spir}|}{S_{rest1}} \cdot 100\% = 5,31\%.$$

IV. CONCLUSION

The development and use of CAD system allows to provide high quality design solutions, significantly reduce the production time. Reducing of the percentage of cuttings can keep production costs down, which generally increases the competitiveness of products.

As a percentage of cuttings in the case of layout of a right circular truncated cone in the form of flat patterns is more than 10%, and for the case of the layout of the spiral scan, the percentage of cuttings is 5.5%, the proposed method of the layout of flat patterns of the surface of a cone provides a significant saving of materials on which it is made.

The proposed method of the layout of spiral scans can be used for various materials of light industry, as well as for the manufacture of scans of conical surfaces of metals, plastics and other materials.

ACKNOWLEDGMENT

The research has been carried out at the expense of the Grant of the President of the Russian Federation for state support of young Russian scientists (MK-3403.2018.8).

REFERENCES

- [1] R. L. Carrel, "Experimental investigation of conical spiral antenna", University of Illinois, Antenna Laboratory technical report no. 22, (1957).
- [2] A. Jiwani and S. Padhi, "Modelling and Simulation of Conical Spiral Antennas", University of Cambridge, UK, AAVP workshop, (2010).
- [3] R. Zhan and X. Adayi, "Formational principle and accurate fitting methodology for a new tooth surface of the spiral bevel gear", Xinjiang University, China, Int. J. Simulation and Process Modelling, vol. 11, no. 1, (2016),
- [4] A. Socas, "Graphics Programming for Designing Conical-Helix Surface", Vilnius Gediminas Technical University, Lithuania, (2015),
- [5] L. A. Bekmurzaev. Reduction of material and labor costs for production: Study guide – Shakhty: Publishing House of SRTUES, 2004. 155 p.
- [6] L. A. Bekmurzaev. Development of a new approach to the design of exclusive clothing models / Bekmurzaev L. A., Byrdina M. V., Nazarenko E. V. // Clothing industry. – 2014. No. 3. P. 24-26.
- [7] G. M. Androsova, "Optimizing of selection of a square of cloth from the matrix elements for a range of fur and leather products", Omsk State Institute of Service, Omsk, (2011).
- [8] The main provisions on the organization of rationing, accounting and rational use of materials in enterprises that produce garments for individual orders of the population. M.:CBNTI, 1994.
- [9] A. Fox and M. Pratt, "Computational geometry", Application in design and manufacturing, M.: Mir, 1982.
- [10] E. H. Melikov. Laboratory course in technology of garments – M.: Legprombytizdat, 1988. 272 p.
- [11] G. I. Surikova, Pattern layouts of clothing parts in CAD: Study guide – Ivanovo: IGTA, 2005. 152 p.
- [12] The basic conditions on the organization of yardage, accounting and rational use of materials at the enterprises that manufacture garments for individual orders of the population. M.: CBNTI, 1994.
- [13] N. N. Razdomakhin. Analytical description of involutes of the tridimensional surfaces of the mannequin and garment, Garment industry, no. 6, (1997), pp. 35.
- [14] Norenkov, I. P. Fundamentals of computer-aided design: Textbook for universities / I. P. Norenkov. – M.: Publishing House of MGTU named after Bauman, 2002. 336 p.

??

Secure Communication using the synchronization of time-varying complex networks by fuzzy impulsive method

Reza Behinfaraz¹, Sehraneh Ghaemi², Sohrab Khanmohammadi³ and Mohammad ali Badamchizadeh⁴

Abstract—In this paper synchronization problem of time-varying complex networks has been studied. Based on impulsive control analysis, a new controller is obtained. Simulation example shows the effectiveness of the proposed method for the time varying complex networks and secure communication using this method is investigated.

Index Terms—Synchronization , Complex network , Impulsive control , Secure communication

I. INTRODUCTION

According to the large variety of complex networks in the different fields such as social systems [1], physical system [1] and biological systems [2], attention to these networks becomes more everyday. In many real world application of complex networks, failure of a link between two nodes of network or creation of a new link is usual, then study on this types of complex networks seems to be important. Time varying complex systems are considered in some studies [3], [4]. But to the best of our knowledge the case that these changes don't allow any specific order is a new topic on complex systems with switching topologies. . Most of the studies on the complex networks have been done on the fixed networks with constant links between nodes of networks. Switching topology for the complex network is a topic which is introduced in recent years [5], [6]. In real world applications switching topology for the complex network is a common topic which happens in many networks. Some researches have been done for synchronization in this area [7], [8]. Synchronization of complex networks is a challenging problem which has wide variety of application in different fields [9], [10], [11], [12], [13]. Different types have been defined for the synchronization problem such as projective [14] In this paper we propose a new method for secure communication of information signals using the synchronization of chaotic networks. The proposed method is obtained using the fuzzy approach which has the good performance with time varying networks.

II. BASIC DEFINITIONS

A. Switching complex network

A switching network with N coupled identical nodes, is considered as

$$\dot{x}_i(t) = f(x_i(t)) - c \sum_{j=1}^N L_{ij} x_j(t) \quad (1)$$

Where $i = 1, 2, \dots, N$ and N is the number of network nodes, $x_i(t) \in R^n$ is de dynamics of each node of network. $f : \in R^n \rightarrow R^n$ is the nonlinear vector function in each node. $c \in$

R is coupling coefficient and $L_{ij}(t)$ is a function satisfying $L_{ij} = 0$ or $L_{ij} = 1$ in each time interval $[t_k, t_{k+1})$.

B. T-S fuzzy system

Some if-then rules together are formed a fuzzy model. Modeling of nonlinear system using T-S fuzzy model is done by local linearization. These local linear subspace become with each other to generate the main nonlinear system model. Consider a general version of a network as Eq.(1) then a T-S fuzzy model for this network can be represent as:

Rule j : if $z_1(t)$ is M_{j1} and $z_2(t)$ is M_{j2} and ... and $z_p(t)$ is M_{jp} , then

$$\dot{x}_i(t) = A_j x_i(t) - c \sum_{j=1}^N L_{ij} x_j(t) \quad (2)$$

where $z_1(t), \dots, z_p(t)$ are new variables and the fuzzy sets are shown by M_{ij} . Also A_i is constant matrix and number of fuzzy rules is shown by r .

The output of fuzzy model (2) is calculated using the singleton fuzzifier, product fuzzy inference and weighted average defuzzifier as:

$$\dot{x}_i(t) = \sum_{i=1}^r h_j(z(t))(A_j x_i(t)) - c \sum_{j=1}^N L_{ij} x_j(t) \quad (3)$$

where $z(t) = (z_1(t), \dots, z_p(t))$ and

$$h_i(z(t)) = \frac{w_i(z(t))}{\sum_{i=1}^r (w_i(z(t)))}, w_j(z(t)) = \prod M_k^j(z_k(t)) \quad (4)$$

III. FUZZY IMPULSIVE CONTROL

In this part, design of a fuzzy impulsive controller is discussed. According to te described structure the fuzzy control rules are defined as the follows

Rule j : if $z_1(t)$ is M_{j1} , $z_2(t)$ is $M_{j2}, \dots, z_p(t)$ is M_{jp} , then

$$u_j(t) = \sum_{m=0}^{\infty} \delta(t - \tau_m) J_{jm}(X(t)) \quad (5)$$

where $\delta(t)$ denotes the Dirac delta function, also the additive change of the state at time k is shown by $J_{jm}(x)$ and $J_{jm}(0) = 0$. Also the sequence of time as τ_m satisfy $0 < \tau_1 < \tau_2 < \dots < \tau_m < \tau_{m+1}$, $\lim_{t \rightarrow \infty} \tau_m = \infty$.

IV. SECURE COMMUNICATION

A. Transmitter

For a transmitter, a network with N nodes as Eq.1 without coupling can be written as the follows.

$$\dot{x} = f_i(x_i(t), a_i(t)) \quad (6)$$

where $x_i(t) = [x_1(t), x_2(t), \dots, x_n(t)]^T$ is the state vector, $a_i(t) = [a_1(t), a_2(t), \dots, a_r(t)]$ shows the parameters vector with embedded messages. Using the chaotic parameter modulation, parameters of network(1) is transferred using the message signals as $s(t) = [s_1(t), s_2(t), \dots, s_r(t)]^T$, by chaotic parameter modulation. The modulation rule for modulate message signal in parameter of system is selected as:

$$a_i(t) = a_i + s_i(t) \quad (7)$$

where a_i is a constant vector.

B. Receiver

The second network is the receiver network. Each node of this network is described as:

$$\dot{x}_i^r = f_i(x_i^r(t), a_i^r(t)) \quad (8)$$

$x_i^r(t) = [x_1^r(t), x_2^r(t), \dots, x_n^r(t)]^T$ shows the state vector of this network, u_i is the control input and a_i^r denotes the parameters of each node. Also s_i^r shows the message signal which is recovered in receiver. The modulation in the receiver is formulated as:

$$a_i^r(t) = a_i^r + s_i^r(t) \quad (9)$$

Now the main problem is the finding of appropriate u_i, a_i^r and s_i^r such that the (8) follows the trajectories transmitter (6) of receiver (8) converges to the transmitter (6). In other expression, the problem is to find appropriate u_i, a_i^r and s_i^r of network (8) satisfy $\|x_i - x_i^r\| \rightarrow 0, \|a_i - a_i^r\| \rightarrow 0$ and $\|s_i - s_i^r\| \rightarrow 0$, with $t \rightarrow \infty$.

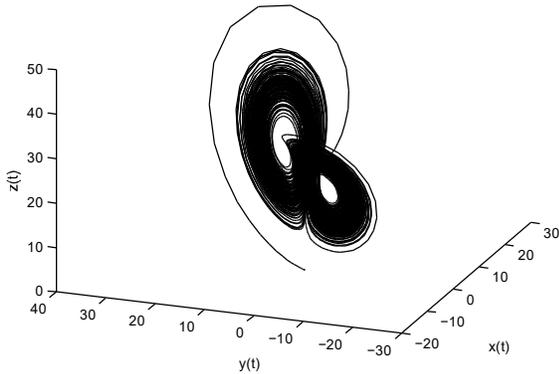


Fig. 1. Chaotic attractor of unified system with varying parameter.

C. Synchronization

In this section, synchronization problem between tow networks are discussed. The dynamic of each node in each networks is considered as unified chaotic system [15]. Dynamic of each node (with m nodes) $i, 1 \leq i \leq m$ is formulated by

$$\begin{cases} \dot{x}_{i1} = a_{i1}(x_{2i} - x_{i1}) \\ \dot{x}_{i2} = a_{i2}x_{i1} + a_{i3}x_{i2} - 10x_{i1}x_{i3} \\ \dot{x}_{i3} = 10x_{i1}x_{i2} - 3x_{i3} \end{cases} \quad (10)$$

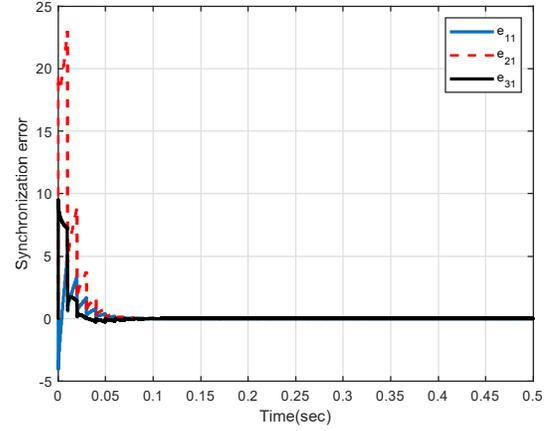


Fig. 2. Errors of first node synchronization between two different complex networks

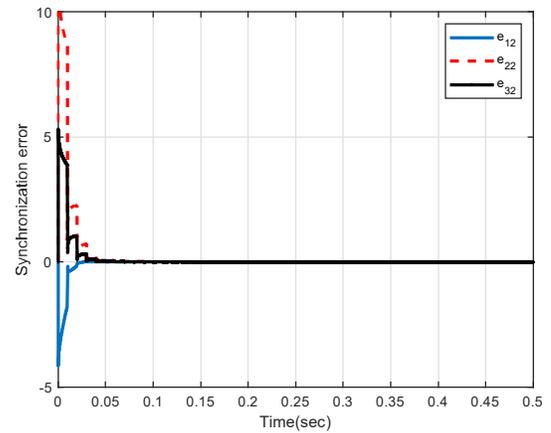


Fig. 3. Errors of second node synchronization between two different complex networks

The parameter selection are as $a_{i1} = 35, a_{i2} = 7, a_{i3} = 28$, where with this parameter system (10) has a chaotic behavior. The second network is defined as the follows:

$$\begin{cases} \dot{x}_{i1}^r = a_{i1}^r(x_{2i}^r - x_{i1}^r) + u_{i1} \\ \dot{x}_{i2}^r = a_{i2}^r x_{i1}^r + a_{i3}^r x_{i2}^r - 10x_{i1}^r x_{i3}^r + u_{i2} \\ \dot{x}_{i3}^r = 10x_{i1}^r x_{i2}^r - 3x_{i3}^r + u_{i3} \end{cases} \quad (11)$$

The message signals are as $s_i = [s_{i1}, s_{i2}, s_{i3}]$. These signals are modulated with system parameter as: are given by

$$\begin{cases} a_{i1}(t) = a_{i1} + s_{i1}(t) \\ a_{i2}(t) = a_{i2} + s_{i2}(t) \\ a_{i3}(t) = a_{i3} + s_{i3}(t) \end{cases} \quad (12)$$

This modulation for the reviver side is as:

$$\begin{cases} a_{i1}^r(t) = a_{i1}^r + s_{i1}^r(t) \\ a_{i2}^r(t) = a_{i2}^r + s_{i2}^r(t) \\ a_{i3}^r(t) = a_{i3}^r + s_{i3}^r(t) \end{cases} \quad (13)$$

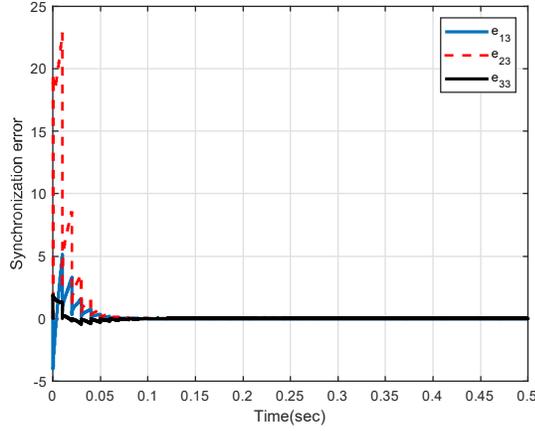


Fig. 4. Errors of third node synchronization between two different complex networks

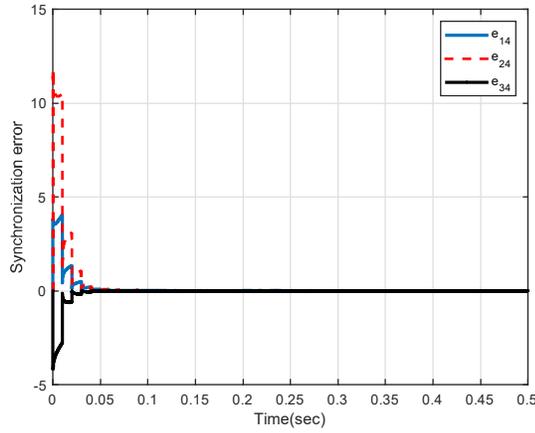


Fig. 5. Errors of fourth node synchronization between two different complex networks

The error signals are defined as:

$$\begin{cases} e_{i1}(t) = x_{i1}^r - x_{i1}(t), \bar{a}_{i1}(t) = a_{i1}^r - a_{i1}, \bar{s}_{i1}(t) = s_{i1}^r - s_{i1}(t) \\ e_{i2}(t) = x_{i2}^r - x_{i2}(t), \bar{a}_{i2}(t) = a_{i2}^r - a_{i2}, \bar{s}_{i2}(t) = s_{i2}^r - s_{i2}(t) \\ e_{i3}(t) = x_{i3}^r - x_{i3}(t), \bar{a}_{i3}(t) = a_{i3}^r - a_{i3}, \bar{s}_{i3}(t) = s_{i3}^r - s_{i3}(t) \end{cases} \quad (14)$$

Where \bar{a}_{ij} and \bar{s}_{ij} for $j = 1, 2, 3$ are parameters differentiation and error of information signal, respectively. Final form of synchronization can be represented as:

$$\begin{cases} \dot{e}_{i1}(t) = (\bar{a}_{i1} + \bar{s}_{i1})(e_{i1} - e_{i1}) - u_{i1} \\ \dot{e}_{i2}(t) = (\bar{a}_{i2} + \bar{s}_{i2})e_{i1} + (\bar{a}_{i3} + \bar{s}_{i3})e_{i2} - 10x_{i1}e_{i13} - u_{i2} \\ \dot{e}_{i3}(t) = -3e_{i3} - u_{i3} \end{cases} \quad (15)$$

V. NUMERICAL SIMULATION

In this section effectiveness of the proposed method for synchronization of two different complex networks is presented by simulation example. Two 4-node complex networks with unified chaotic system are considered. Variation on the

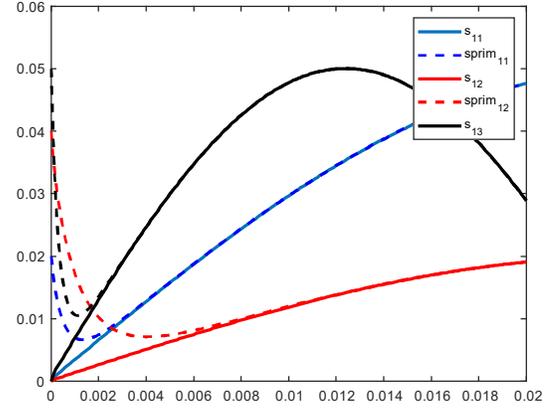


Fig. 6. Information signals in transmitter and receiver

parameters of unified system are considered. Fig.1 shows the behavior of chaotic unified system. The information signals s_{i1} , s_{i2} and s_{i3} are selected as follows:

$$\begin{cases} s_{i1}(t) = 0.05\sin(200\pi t), \\ s_{i2}(t) = 0.02\sin(200\pi t), \\ s_{i3}(t) = 0.05\sin(400\pi t), \end{cases} \quad (16)$$

for $i = 1, 2, 3$. Synchronization errors for each node of network are shown in Figs 2, 3,4, 5. Also information signals in transmitter and receiver for node 1 is shown in Fig.(6) for the other nodes these results are the same.

VI. CONCLUSION

In this paper, secure communication using the synchronization of chaotic networks was proposed. The control approaches were based on fuzzy logic. It was shown that this controller has ability in the synchronization of time varying network. Results were showed the performance of the proposed method.

REFERENCES

- [1] Pastor-Satorras R, Vespignani A. Epidemic dynamics and endemic states in complex networks. *Physical Review E*. 2001 May 22;63(6):066117.
- [2] Moreno Y, Pastor-Satorras R, Vespignani A. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems*. 2002 Apr 1;26(4):521-9.
- [3] Lu J, Chen G. A time-varying complex dynamical network model and its controlled synchronization criteria. *IEEE Transactions on Automatic Control*. 2005 Jun;50(6):841-6.
- [4] Tang J, Scellato S, Musolesi M, Mascolo C, Latora V. Small-world behavior in time-varying graphs. *Physical Review E*. 2010 May 17;81(5):055101.
- [5] Cui W, Fang JA, Zhang W, Wang X. Finite-time cluster synchronization of Markovian switching complex networks with stochastic perturbations. *IET Control Theory & Applications*. 2014 Jan 2;8(1):30-41.
- [6] Liu T, Zhao J, Hill DJ. Exponential synchronization of complex delayed dynamical networks with switching topology. *IEEE Transactions on Circuits and Systems I: Regular Papers*. 2010 Nov;57(11):2967-80.
- [7] Ali MS, Yogambigai J, Jinde CA. Synchronization of master-slave Markovian switching complex dynamical networks with time-varying delays in nonlinear function via sliding mode control. *Acta Mathematica Scientia*. 2017 Mar 1;37(2):368-84.

- [8] Reza Behinfaraz, Mohammadali Badamchizadeh, Amir Rikhtegar Ghiasi; An adaptive method to parameter identification and synchronization of fractional-order chaotic systems with parameter uncertainty; *Applied Mathematical Modelling*; 40 (2016) 4468-4479.
- [9] Reza Behinfaraz, Mohammadali Badamchizadeh; Optimal synchronization of two different incommensurate fractional-order chaotic systems with fractional cost function; *Complexity*; 21(2016) 401-416.
- [10] Reza Behinfaraz, Mohammad Ali Badamchizadeh; Synchronization of different fractional-ordered chaotic systems using optimized active control; 6th International IEEE Conference on Modeling, Simulation, and Applied Optimization (ICMSAO), 2015 ; 2015: 1-6.
- [11] Behinfaraz R, Ghaemi S, Khanmohammadi S. Risk assessment in control of fractional-order coronary artery system in the presence of external disturbance with different proposed controllers. *Applied Soft Computing*. 2019 Apr 1;77:290-9.
- [12] Behinfaraz R, Ghaemi S, Khanmohammadi S. Adaptive synchronization of new fractional-order chaotic systems with fractional adaption laws based on risk analysis. *Mathematical Methods in the Applied Sciences*. 2019 Apr;42(6):1772-85.
- [13] Behinfaraz R, Badamchizadeh MA. New approach to synchronization of two different fractional-order chaotic systems. *IEEE International Symposium on Artificial Intelligence and Signal Processing (AISP) 2015 Mar 3* (pp. 149-153).
- [14] Behinfaraz R, Badamchizadeh MA, Ghiasi AR. An approach to achieve modified projective synchronization between different types of fractional-order chaotic systems with time-varying delays. *Chaos, Solitons & Fractals*. 2015 Sep 1;78:95-106.
- [15] Behinfaraz R, Badamchizadeh MA. Synchronization of different fractional order chaotic systems with time-varying parameter and orders. *ISA transactions*. 2018 Sep 1;80:399-410.

Description of the spatial shape surface of an air supported dynamic figure

Mikhail F. Mitsik
Mathematics and applied informatics
Don State Technical University
Rostov-on-Don, Russia
m_mits@mail.ru

Lema A. Bekmurzaev
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
Bekmurzaev.l@yandex.ru

Marina V. Byrdina
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
byrdinamarina@mail.ru

Olga A. Aleynikova
Mathematics and applied informatics
Don State Technical University,
Rostov-on-Don, Russia
aleynikova.o@mail.ru

Victor N. Kokhanenko
General engineering disciplines
Platov South-Russian State Polytechnic University(NPI)
Novocherkassk, Russia
d-kelekhsaev@mail.ru

Abstract— The paper proposes the analytical method of design of structures in the form of a flexible inextensible shell. Research topic of this paper corresponds to scientific directions of symposium System-in-Package and 3D Design & Test. As an example of such a design, a product is to promote the development of microelectronics - an aerodynamic figure is considered, which is a connected set of unfolding surfaces. The aero-figure is a 3D object and is designed in Maple based on the package commands with 4GL library. Analytical construction of 3D graphical objects is implemented on the basis of commands that allow to draw surfaces and spatial curves specified in explicit and parametric form. All elements of the aero-figure are unfolding surfaces of conical type and intersect with each other only on common boundaries – space curves. The current approach allows to design an aero-figure of specified dimensions based on the calculation of the scan elements.

Keywords— computer simulation, aerodynamic figure, 3D graphic objects, 2D scans, package Maple 2015

I. INTRODUCTION

Modeling and computer-aided design have a very significant impact on all practical human activities. Appearance of CAD instruments – computer tools for microelectronic design engineering was caused by rapidly increasing of complexity and amount of tasks. Optimal placement and connection of functional modules of complex functional devices [1] doesn't seem to be possible without using CAD tools anymore.

Requirements for integral schemes design instruments are continuously increasing and this is why semiconductor devices manufacturers forced to step up efforts to develop new tools. Because of there are being developed high-level logical modeling and synthesis systems used for creating of integral schemes and digital technologies universal integral design environment. Experts of Synopsys and Cadence entered the market of automatic synthesis based on standard logic element libraries.

In 21st century, production began using technological standards of 130, 90 and even 50 nm. First Encounter system became standard for nanometer technologies design. One of ways to promote new design technologies is creating of aerodynamic figure which is shell construction [2]. Shell construction might have multi-layer design with each layer

having its own technological purpose and layer width might be designed with nanometer technologies.

Dynamic air-supported figure is a creative type of advertising that attracts attention with its unusual shape and large size. Aero figure is projected from 3 to 10 meters high. A dancing inflatable person with an aesthetic appearance is visible from afar, it pleases both adults and children, and therefore, it is suitable for both advertising purposes and promotion of microelectronics products. A dancing inflatable person can be installed near the shopping center or store, at the entrance of the cafe, the exhibition, amusement park, during a variety of carnivals and festivals.

On this subject at present, there are only patents for invention [3] and for the utility model [4]. In [3], a description of the work of an aero figure at constant or variable air flow into its legs is proposed; various options for air outlets are considered that can significantly change and diversify the set of dance movements of the aero figure. An explanation is given why the air figure under the influence of the air flow makes wave-like cyclic movements. The description of specific models of fans and products with specific dimensions is given. The conditions of safe operation of the air figure are described, a qualitative description of the geometry of the air figure for its reliable operation is pointed out.

In the paper [4], a utility model was described, which consists in creating a movable air-supported structure that simulates a living creature on its head and performs various movements associated with acrobatic or dancing movements. The hole of the sleeve, imitating the head, is connected to the output of the airflow generator. In this design, there is a "single-point" hinged fastening of the shell to the generator of the air flow, which is less stable than the two-point one. At the same time, the sleeve of the shell simulating the head works not only for bending, but also for torsion, to extend the range of movements of the shell elements of this tool during its demonstration.

The purpose of the paper is to create an analytical method of designing deployable air supported dynamic figure of given geometric dimensions, to describe the volume-spatial form of the product prototype using application package.

Research tasks are: 1) to describe the shape of the elements that make up the air supported figure, to describe the curves by which the elements of air supported figure are interconnected; 2) to improve a traditional form of an aero figure for the convenience of its practical application; 3) to make "assembly" of an aerodynamic figure from its constituent elements; create a program for the automatic design of aero figures.

The novelty of the work is: 1) in the analytical description of the surface of the aerodynamic figure and its constituent elements in Cartesian rectangular coordinates and in the parametric coordinate system; 2) in creating a program for building supported figure depending on the given geometrical dimensions; 3) in the possibility of promptly changing the shape of the product, the color of each elements of the aero figure according to the customer's taste and demonstration long time before creating the product.

Computer modeling of physical and geometric similarity of a prototype to a real object is necessary in order to study the properties of a technical object, its visualization long before the creation of the object itself [5].

Research topic of this paper corresponds to scientific directions of symposium 1) System-in-Package and 3D Design & Test, 2) CAD and EDA Tools, Methods and Algorithms.

II. SURFACE OF AEROFIGURES IN MAPLE

The Maple 2015 package has interactive graphics capabilities and allows to visualize products based on symbolic mathematics [6]. Maple has a very user-friendly programming language that has a crisp logic that is clear not only to a professional programmer. In addition, Maple allows you to create simple application packages and integrate them into a complex program. It automatically selects the required types of variables and checks the correctness of operations, does not require a description of variables and strict formalization of the record. Graphic and spreadsheet management simplifies the user interaction with the Maple engine, performing the role of a tool, with the help of which the requests to perform specific tasks are passed and the results are displayed.

The Maple system supports two-dimensional and three-dimensional graphics, it makes possible to combine graphics of several functions, to represent explicit, implicit and parametric dependencies, as well as multidimensional functions and simply a set of data in a graphical form and visually determine the geometry of objects. Maple builds surfaces and curves in three-dimensional space, including surfaces specified in explicit, implicit, or parametric forms [7], and visualizes solutions to differential equations. In this case, graphical objects can be represented not only in static form, but also in the form of two-dimensional or three-dimensional animation. This feature of the system can be used to display real-time processes.

The "body" of the aero-figure is a conical surface, which is bounded from above by a cone connecting the "shoulders" of the aero-figure with its "neck". From below, the "body" of the aero-figure is bounded to the "legs", each of them being also a circular cone. In this case, the axis of symmetry of the "body" coincides with the vertical axis, and each of the axes for the "legs" is inclined to the vertical axis

at an acute angle (Fig. 1). The lower border of the body, marked in Fig. 1 in red, does not lie in the horizontal plane, since it is the intersection line between the "body" and "legs".

The equation of the surface of the "body" of the aero-figure in Cartesian rectangular coordinates is the following

$$x^2 + y^2 = p^2 \cdot (z - a)^2, \quad (1)$$

where x, y – Cartesian coordinates in the horizontal plane,

z – vertical coordinate;

$p = \text{arctg } \alpha$, α – angle between the axis of rotation of the cone and its generator;

a – magnitude of the shift in the direction of the OZ axis.

In the paper the parameter p is accepted equal to $p = \frac{1}{12}$.

In Maple, the surface is better drawn if it is defined in cylindrical coordinates system

$$r = p \cdot (z - a), \quad (2)$$

where r – polar radius in the horizontal plane.

The surfaces representing the "legs" of the aero-figure are also cones of rotation, but they are inclined to the axis of the applique at an acute angle β , which is set by the designer himself, in the paper it is taken equal to $\beta = 2\pi/5$. Accordingly, to vector the inclined cones of rotation, it is necessary to perform the transformation of the rotation of the conical surface in the XOZ-plane. The formulas for the rotation transformation in the XOZ-plane are the following

$$\begin{cases} X = x \cos \beta + z \sin \beta; \\ Y = y; \\ Z = -x \sin \beta + z \cos \beta. \end{cases} \quad (3)$$

In this regard, the analytical description of the "legs" of the aero-figure in the Cartesian coordinate system is inconvenient. In the cylindrical coordinate system, the surface equations describing the "left leg" of the aero-figure were calculated in the formula

$$\begin{cases} X = 3 \cos t \cdot |z/(3p) - 1|; \\ Y = 3 \sin t \cos \beta + z \sin \beta; \\ Z = -3 \sin t \sin \beta \cdot |z/(3p) - 1| + z \cos \beta + 4. \end{cases} \quad (4)$$

Accordingly, dependencies were obtained for the right "leg" of the aero-figure

$$\begin{cases} X = 3 \cos t \cdot |z/(3p) - 1|; \\ Y = 3 \sin t \cos \beta + z \sin \beta; \\ Z = 3 \sin t \sin \beta \cdot |z/(3p) - 1| - z \cos \beta - 4. \end{cases} \quad (5)$$

The curves that are the boundaries of the "body" and simultaneously the "legs" of the aero-figure were obtained, the lines of intersection of surfaces (2) and (4), (2) and (5), respectively. In addition, the surfaces (4) and (5) intersect.

The elements “body”, “legs” and lines of their intersection are shown in Fig. 1.

Since the surfaces (4) and (5) are symmetrical with respect to the vertical axis, the intersection curve highlighted in red is in the vertical plane.

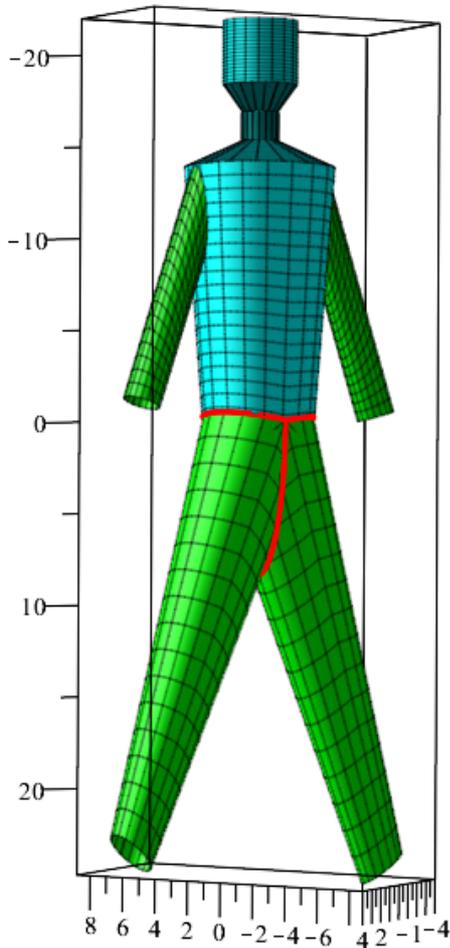


Fig. 1. Design of the aero-figure surface in Maple

The equation of the intersection curve of surfaces (4) and (5) is the following

$$\begin{cases} X = 3 \cos t + z \sin \beta; \\ Y = 0; \\ Z = \frac{0,85 \cdot 3 \cdot \sin t}{\cos \beta}. \end{cases} \quad (6)$$

The equations of the boundary between the elements of the “body” and the “left leg”, respectively, the “right leg” differ only in the sign of the Y coordinate: minus sign to describe the border with the “left leg” and plus sign for the border with the “right leg”

$$\begin{cases} X = 3 \cos t; \\ Y = \pm 3 \sin t; \\ Z = 3 \sin t \frac{1 + \sin \beta}{\cos \beta}. \end{cases} \quad (7)$$

The surfaces imitating the “hands” of the aero-figure are chosen in the form of circular cylinders, in which each of the edges bordering the “body” is a closed space curve. The equations describing the cylindrical surfaces of the “hand” taking into account the rotation transformations, as well as their some distance from the “body” for support view are selected in the form

$$\begin{cases} X = \cos t; \\ Y = \pm(\sin t \cdot \sin \beta + z \cos \beta + 13,3); \\ Z = \sin t \cdot \cos \beta - z \sin \beta + 8,7, \end{cases} \quad (8)$$

where $z = 10 \dots 21 - 2,5 \sin t, t = 0 \dots 2 \cdot \pi$.

The cones connecting the “body” with the “neck” and the “neck” with the “head” are straight circular and truncated with the OZ axis of rotation, their equations in the cylindrical coordinate system have the form (2) with the corresponding values of the parameters p and a .

In the papers [4, 5] mathematical models of the aerodynamic figure were proposed, but the geometry of the proposed models had several disadvantages (Fig. 1). The surface of the aero-figure can be considered as the first approximation to the surface of the human body [8, 9].

III. MODIFIED SURFACE OF AEROFIGURES

However, in contrast to the trousers, in the proposed aero-figure, the distance of the crotch depth (i.e. the distance from the waist line to the hip line) was much greater than that of a person. At the same time, the width of the legs of the aero-figure was not narrowed to the bottom and its legs were straddled. The geometry of the body was also approximated to the surface of the human body. The purpose of this paper is to eliminate the previously existing shortcomings in the modeling of aero-figure.

Maple uses procedural fourth generation language (4GL), which is designed for rapid development and implementation of mathematics programs and applications for users. The mathematical model of the aerodynamic figure was developed in Maple, with its help the code in C related to this model was generated.

The 4GL language is specially optimized for the development of mathematical applications, it allows to speed up the development process, and the user interface is configured with Maplets elements and Maple documents with built-in graphical components. It is also possible to create interactive documents and presentations in Maple by adding buttons, sliders, and other components, as well as by publishing materials online and deploying interactive computing on the network using the MapleNet server.

The aerodynamic shape is a product of a flexible non-stretchable fabric, inside of which the air is pumped. The air is pumped from below, through the legs of the aero-figure with the fans. The lower border of each leg is tightly fixed at the outlet of the fan and the air moves vertically upwards inside the body of the aero-figure, while inside the aero-figure, a pressure is greater than the atmospheric pressure. This effect ensures the formation of the product. The neck of the aero-figure has a hole (Fig. 2) of a smaller radius than the body that allows it to keep its shape, then the neck merges into the head, on the border which has adjustable hole (exhaust). There is a free flow of air through the holes

in the head and hands of aero-figure. Due to the pulsating nature of the distribution of air velocities and pressures inside the aero-figure, on the one hand, it “keeps its shape”, and on the other hand – makes dance movements. An advanced model of the aerodynamic figure, made in the Maple 2015, in the construction of which the above shortcomings were eliminated, is presented in figure 2.

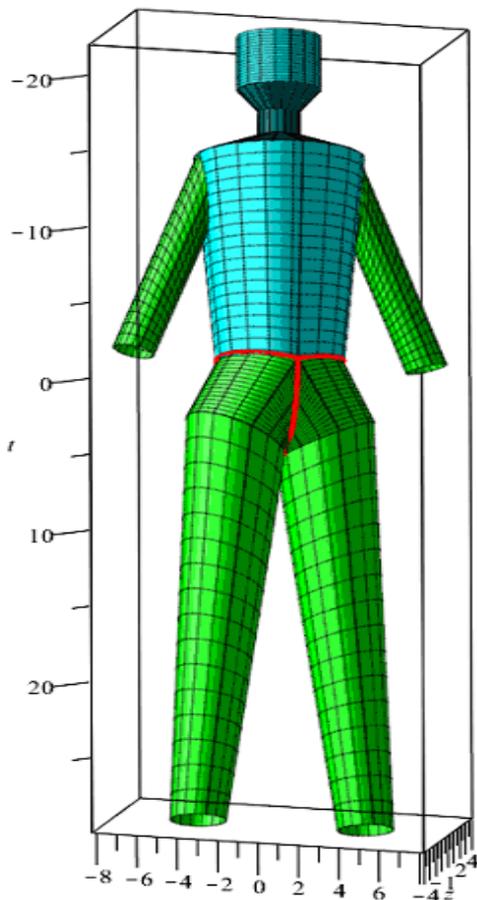


Fig. 2. Advanced model of the aero-figure surface in Maple

Aero-figure can be considered as a flexible inextensible shell [10], its study has theoretical and practical value, as a design that does not obey the laws of pneumostatics [11]. It is a self-vibrating object, the shape of which depends significantly on its mass, the moment of inertia and the air flow generator. The calculation of velocity and pressure distribution in this design are made with non-linear dependencies on the basis of the equations of motion for viscous gas, Navier – Stokes equations.

The enveloping surfaces are surfaces of zero Gaussian curvature, their advantage lies in the possibility of their deployment and overlaying these surfaces on the plane by means of bending.

IV. CONCLUSION

The surface of the aerodynamic figure is a set of elements of the enveloping surfaces, that greatly simplifies its design. The construction of two-dimensional scans of the elements of the aero-figure significantly saves materials for its making as well as labor costs.

On the basis of the developed program for the design of aerodynamic figures, it is possible to develop an automated system that will reduce the material and labor costs for making new products by creating virtual samples of 3D models, the possibility of rapid changes in the model samples and their transmission over the Internet. The program can be useful in the design of elements of shell materials, as well as for textiles designers.

Result of this paper could be used in System-in-Package and 3D Design & Test, CAD and EDA Tools, Methods and Algorithms scientific directions.

ACKNOWLEDGMENT

The research has been carried out at the expense of the Grant of the President of the Russian Federation for state support of young Russian scientists (MK-3403.2018.8).

REFERENCES

- [1] H. Lehmann, S. Menzel. Evolvability as concept for the optimal design of free-form deformation control volumes. 2012 IEEE Congress on Evolutionary Computation.
- [2] M.F Mitsik, M.V. Byrdina, L.A. Bekmurzaev. Modeling of developable surfaces of three-dimensional geometric objects. Proceedings of 2017 IEEE East-West Design and Test Symposium, EWDTS 2017 2017. C. 8110086.
- [3] Patent USA 6186857 B1, Apparatus and method for providing inflated undulating figures, Doron Gazit, Arieh Leon Dranger, Los Angeles, 2001.
- [4] The patent for utility model No. RU 70724 Russian Federation, IPC U1 G09F 19/08 Movable air-supporting structure / A.N. Komissarov; the applicant and patent holder of NPO Aero ecology LLC. No. 20071239396/22 appl. August 1, 2007; publ. 10.02.2008 Bull. No. 45.
- [5] Mitsik M.F., Movchun A.A. Two-Dimensional and three-dimensional visualization of products in the Maple / Scientific and technical Bulletin of the Volga region. 2018. No. 1. P. 132-135.
- [6] Official website of Maple [Electronic resource]. – 2018. Available at: <https://www.maplesoft.com/products/maple/professional/>
- [7] M.V. Kiseleva. Development of parametric method of 3D modeling of women’s lower body garments: Ph.D. thesis in Engineering Science: 05.19.04 /MGUDT, Moscow, 2011. – 232 p.
- [8] V.D. Frolovsky. Modeling of tissue behavior on the surface of a computer dummy // News of Higher Educational Institutions, technology of textile industry – 2006. – №4 –P. 68-71.
- [9] N.N. Razdomakhin. Analytical description of volumetric scans of the surfaces of the mannequin and clothing // Sewing industry, 1997. – №6. – P. 35.
- [10] M.V. Byrdina. Design of exclusive models of clothes with the use of analytical method of scanning // Clothing industry. – 2014. - № 3. – P. 40-41.
- [11] Sze K. Y., Liu X. H. A corporation grid-based model for fabric drapes // Int. J. Numer. Methods Eng., 57 – 2003. – P. 1503-1521..

??

Solution of the Dynamic Problem for Optimal Design of Electronic Devices Based On the Gravity Center Method

M. V. Donadze¹

Department of Computer Sciences
¹Batumi Shota Rustaveli State University
 Batumi, Georgia
mikheil.donadze@bsu.edu.ge

Z. Meskhidze²

PHD studies
²Georgian Technical University
 Tbilisi, Georgia
zurab.meskhidze@gmail.com

Abstract - Based on the gravity center methodology algorithm is designed for solving dynamic problems of parametric optimization for electrical circuits. The approach used in the algorithm enables the derivation of infinite dimension problems into finite dimension ones and solving wide spectra of optimization problems with the persistence tolerable in engineering practice and minimal computational power cost.

Key words: Transistor Amplifier, optimization, optimal parameters, design passive components.

I. INTRODUCTION

Increasing interest in designing and managing engineering practices in modern technical systems is essential to the development of effective methods for optimizing the complex (multidimensional, nonlinear, non-referenced, multimodal) tasks. The synthesis of such techniques that provide easy and accurate precision in the automated systems of automated systems with easy, quick and computational time consumption, as well as solving extreme static and dynamic extreme tasks.

One of the main practices in this field is developing numerical methods for solving the optimal management tasks. The optimal management tasks are dynamic tasks for optimization and their solution is to find an extremist functionality [2,3,5,8,9]. Therefore, the mathematical apparatus for solving these types of tasks is a variation. In case of restrictions on condition and management variables, optimal management tasks can be solved based on the principle of Pontragine maximum and Belmann dynamic programming method. These methods, which belong to the gold fund of optimal management theory, have gained widespread popularity due to their generality and well-established theorems, but they are able to solve various tasks with optimal management through only simple cases that are far from the requirements of modern practice.

The principle of maximum is a generalization of the main results of classical variation calculations. It determines the optimal necessary and sufficient conditions for linear dynamic systems, and for non-linear systems only necessary conditions. This means that in non-linear systems it is defined not by optimal management but some small group of admissions, among which there may be an optimal management of the search [4].

Generally, a dynamic programmatic method, as well as its derivative numerical methods, set great requirements for computer memory. It should be noted that in the higher order system than the fourth, where a number of functions are involved, the solution of optimal management tasks through dynamic programmatic method is still a big problem and in many cases it is impossible task [2].

The above mentioned enables us to use the method of gravity centers to solve the optimal management tasks, with the accuracy of the precision and the minimum amount of computer time, can be solved a number of tasks of practical importance, including hard problem solving tasks for optimal design of electronic devices.

II. SOLUTION OF THE DYNAMIC OPTIMIZATION PROBLEMS BASED ON THE GRAVITY CENTER METHOD

Consider the general task of optimal management of nonstandard dynamic system with lumped parameters of continuous action:

$$\min \left\{ J[u(t)] = \int_{t_0}^{t_K} f(y, u, t) dt \mid \dot{y} = \phi(y, u, t), y(t_0) = y_0; y(t_K) = y_K; g(y, u) \leq 0 \right\}. \quad (1)$$

(1) The definition $J[u(t)]$ (and not $J[y(t), u(t)]$) underlines the condition that $u(t)$ is a depended variable while $y(t)$ is seen as solution to the given differential equation.

For solving optimal control (1) task, the following approach is used. From the mathematical perspective, optimal control task is an infinite dimensional problem from mathematical programming in the infinite dimensional plane and for solving which we can only use mathematical programming methods if we derivate this infinite dimensional optimization problem into finite dimensional one [5].

The derivation of the problem (1) to the finite dimensional state is possible by approximation given infinite dimensional functional plane by its finite dimensional sub plane based on which the variable $u(t)$ instead of function, with some approximation can be viewed as linear multivariable with defined structure:

$$u_c(t) \approx u_c(\lambda) = \sum_{i=1}^c \lambda_i \omega_i, (c = 1, 2, \dots), \quad (2)$$

Where λ_i are coefficients that can be defined and ω_i are known analytical functions (e.g. step, polynomial, exponential etc.), which in given plane construct the entire system of functions.

Having in mind relation (2) the optimization function with fixed C is transformed as normal functions of $\lambda_1, \lambda_2, \dots, \lambda_c$ variables

$$J[u_c(t)] = J(\lambda_1, \lambda_2, \dots, \lambda_c), \quad (3)$$

And the optimal control initial problem can be derived as finite dimensional problem from nonlinear programming:

$$\min \left\{ J(\lambda) = \int_{t_0}^{t_K} f(y, u_c, t) dt \mid \dot{y} = \phi(y, u_c, t), y(t_0) = y_0; y(t_K) = y_K; g(y, u_c) \leq 0; \lambda = (\lambda_1, \lambda_2, \dots, \lambda_c) \in Q_\lambda \subset R^c \right\}. \quad (4)$$

It is clear that for a specific case of control function $u_c(t)$, if λ_i quantities are selected so that following condition is satisfied: $\min J(\lambda_1, \lambda_2, \dots, \lambda_c) = J(\lambda_1^*, \lambda_2^*, \dots, \lambda_c^*)$, then the function $u_c^*(t) = \sum_{i=1}^c \lambda_i^* \omega_i$ with the optimal parameters λ_i^* with certain approximation can be viewed as solution for the problem (1).

For solving equation (4) which is the same as to find extremum values of λ_i parameters the of gravity center method is used.

Choice to use center of gravity method for solving optimal control tasks was supported by following factors:

- The gravity center method gives possibility to solve problems from mathematical programming (including concave nonlinear multidimensional programming problems) with tolerable persistence and insignificant computational recourses.
- With the gravity center method using penalty function by considering restrictions and applying random search elements we are able to solve issues related to the two-point limit equations [6].

It is known that the main issue that arises when solving variation limit problems is to simultaneously satisfy large amount of different conditions. Some of such conditions are:

- Minimization of optimization function:

$$J \rightarrow \min; \quad (5)$$

- Satisfaction conditions of limits initial and final state:

$$y(t_0) = y_0, \quad (6)$$

$$y(t_K) = y_K; \quad (7)$$

- Constraints on condition and control variables satisfy the formula:

$$g(y, u_c) \leq 0. \quad (8)$$

One of the ways to avoid those issues is to replace above listed conditions by single but equivalent condition [6].

To avoid final limit condition (7) what is needed to instead of given criteria for optimality is to consider new, generalized criteria, in which the condition of final limit is dealt with. To achieve this we shall introduce a measurement for not satisfying limit conditions:

$$J_1 = \sum_{i=1}^n |y_i(t_K) - y_{Ki}|^x, x = 2; 1; 0.5, \quad (9)$$

Where $y_i(t_K)$ is the i -coordinate of the endpoint of the phase trajectory, y_{Ki} is the i -coordinate of the given end-state. It is clear that the J_1 criterion is a control vector dependent on the $u(t)$ vector, since $y_i(t_K)$, $i = \overline{1, n}$, depends on the control. This function has minimal value when the trajectory endpoint coincides with the given endpoint of the trajectory, with the minimum value being zero. Thus, the task of meeting the final boundary conditions will be reduced to the task of finding the minimum J_1 function.

Having in mind equation (9) new criteria for optimality can be written as this:

$$\bar{J} = J + \gamma_1 J_1 = J + \gamma_1 \sum_{i=1}^n (y_i(t_K) - y_{Ki})^2, \quad (10)$$

Where γ_1 is a significantly large numerical coefficient. On bases of analyzing functional (10) we can conclude by its minimization, not only the minimum value of J will be calculated by also with some approximation the limit conditions (7) will also be satisfied. The accuracy of satisfying limit conditions will be higher with the larger γ_1 coefficient.

Position and control variables (8) failure to restrict the limits of functions may be carried out by means of which the general criterion will be taken into consideration:

$$\begin{aligned} \tilde{J} &= \bar{J} + \gamma_2 \sum_{i=1}^q \left(\frac{g_i(y, u_c) + |g_i(y, u_c)|}{2} \right)^2 = \\ &= J + \gamma_1 \sum_{i=1}^n (y_i(t_K) - y_{Ki})^2 + \gamma_2 \sum_{i=1}^q \left(\frac{g_i(y, u_c) + |g_i(y, u_c)|}{2} \right)^2, \end{aligned} \quad (11)$$

Where γ_2 is a penalty coefficient and q is a quantity of restrictions for which $g_i(y, u_c) > 0$.

In the task of optimal management (1) or the same thing as (4) the $y(t)$ trajectory in the task of solving the system of differential equations is defined (6) in the initial conditions:

$$\begin{cases} \dot{y} = \phi(y, u_c, t), \\ y(t_0) = y_0. \end{cases} \quad (12)$$

(12) The task is a well-known task of the tower and the Runge-Kutta method [8] is used for its numerical integration in the developed algorithm, which provides a high accuracy of the computational process. It is noteworthy that the error of the Rouge-Kutta method is $O(H^5)$, wherein the integrity of $H = t_K - t_{K-1}$.

Thus, considering above noted statements problem (4) can be represented as:

$$\begin{aligned} \min \left\{ \tilde{J}(\lambda) = \int_{t_0}^{t_K} f(y, u_c, t) dt + \gamma_1 \sum_{i=1}^n (y_i(t_K) - y_{Ki})^2 + \right. \\ \left. + \gamma_2 \sum_{i=1}^q \left(\frac{g_i(y, u_c) + |g_i(y, u_c)|}{2} \right)^2 \mid \dot{y} = \phi(y, u_c, t), \right. \\ \left. y(t_0) = y_0; \lambda = (\lambda_1, \lambda_2, \dots, \lambda_c) \in Q_\lambda \subset R^c \right\}. \end{aligned} \quad (13)$$

It must be noted that solving problem (3) with any numerical method requires representing differential equations as well-known equations and integrals as finite summations.

Solving problem (13) using center of gravity method relays on algorithmic assumptions, for which let's introduce following definition:

Definition 2.5.1. For a normal number B let's define set of type:

$$\Omega_\varepsilon(y_K) = \{y \in B \mid \|y - y_K\| \leq \varepsilon\}, \quad (14)$$

Which geometrically represents a sphere with radius ε and center $y_K \in B$ as a ε neighborhood of element y_K which is the same as permitted error neighborhood. The ε neighborhood is defined as (Fig.1):

$$\Omega_\varepsilon = \sum_{i=1}^n (y_i(t_K) - y_{Ki})^2 - \varepsilon^2 \leq 0. \quad (15)$$

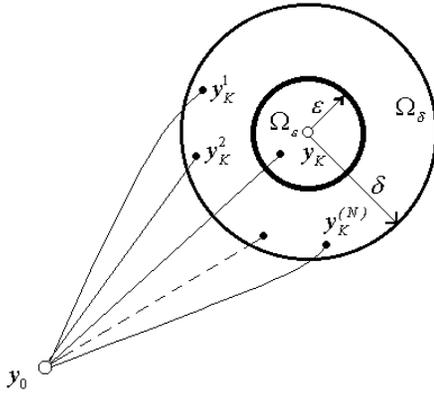


Fig. 1

Definition 2.5.2. For a normal number B let's define set of type:

$$\Omega_\delta(y_K) = \{y \in B \mid \|y - y_K\| \leq \delta; \delta > \varepsilon\}, \quad (16)$$

Which geometrically represents a sphere with radius δ ($\delta > \varepsilon$) and center $y_K \in B$ as δ neighborhood of element y_K which is the same as imaginary neighborhood of the goal. The $\gamma A = \pi r^2 \delta$ neighborhood is defined as (Fig. 1):

$$\Omega_\delta = \sum_{i=1}^n (y_i(t_K) - y_{Ki})^2 - \delta^2 \leq 0. \quad (17)$$

Definition 2.5.3. Let's denote all $y(t)$ phase trajectory that can reach δ neighbor from y_0 point as an acceptable trajectory of controlled process.

Definition 2.5.4. If the acceptable trajectory $y(t)$ of the controlled process reaches ε neighbor lets denote it as acceptable trajectory with accuracy ε .

It is necessary to note that the introduction of the δ - area concepts is related directly to the concepts of control area and control of the dynamic system.

As it is known, the dynamic system is called control if any of the initial y_0 and any final y_K states are controlling $u(t)$ function and finite t_K time that the control of the function $u(t)$ can be transferred from the y_0 to y_K position in the finite t_K time. And the Q_y set of all the y_0 points of the phase space, from which the control goal is to be carried out in the limited time, is called a controlled neighborhood. Obviously, if $y_0 \notin Q_y$, then you can not find a function $u(t)$ function that will transfer the system from y_0 to y_K . Unfortunately, there are no effective methods of defining the control area, in general case.

Only the controllable process of the N number is considered when the optimal solution of the given problem is determined by the centers. In order to achieve a satisfactory result, it is necessary that a broad spectrum of values of the functional functionality in the process of determining the empirical points of $\overline{\lambda_i(p)}$ functions. Obviously, if the neighborhood of control is very limited, then the realization of the different value of the N number of operative functionality is an impossible task (especially in the optimal performance tasks), which makes a significant margin of error.

To avoid it, the concept δ areas was entered that gives the chance artificially to increase area of a final condition of a system and thus to expand the Q_y field of management which will allow us to understand various values of the operating function N .

Increasing the state of the final state of the system will not affect the requirement to meet the boundary conditions, since each deviation from the actual final condition is taken into consideration of the fine function.

Taking into consideration the above, an algorithm for solution of optimal control tasks has been developed, which consists of three stages in structural terms. At the initial stage the values of the Lebesgue level are $\zeta_p, p = \overline{1, K}$, which is a series of preliminary statistical experiments with $S = 0.1N$, where N is the number of major experiments with the following scheme:

1. By random number generator $t \in [t_0, t_K]$ the $t_i = \lambda_i$ random parameters will be generated at the interval and the function $u_c(t)$ function will be based on the formula (2).
2. The solution of the differential equations (12) of the Runge-Kutta method is determined by $y(t)$ trajectory [8].
3. If $y(t)$ is a permeable trajectory, or if $y(t)$ trajectory (17) is an element of magnitude, then (11) the formula is calculated by the generalized \tilde{J} criterion of the optimum, unless otherwise noted in paragraph 1.

In this series of preliminary experiments, the \tilde{J} criterion values are summarized by $J_S = \sum_{j=1}^S \tilde{J}_j$ and minimum value $J_0 = \min_j \{\tilde{J}_j\}$. At the end of the first phase the values of the Lebesgue levels are calculated by the following formula:

$$\zeta_p = \frac{1}{S} J_S - (p - 1) \Delta \zeta, p = \overline{1, K}, \quad (18)$$

Where,

$$\Delta \zeta = \frac{1}{\rho} \left(\frac{1}{S} J_S - J_0 \right). \quad (19)$$

(19) The ρ coefficient in the image is characterized by the density of the Lebesgue levels. Based on practical considerations, we can get $\rho = 10, 15$ or 20 .

The second phase is a series of basic statistical experiments with the number of empirical points of the $\overline{\lambda_i(\zeta)}$, $i = \overline{1, c}$ functions based on the following formula:

$$\overline{\lambda(\zeta_p)} = \frac{\sum_{j=1}^L \lambda^{(j)} (j(\lambda^{(j)} - \zeta_p) \theta(\lambda^{(j)}, \zeta_p))}{\sum_{j=1}^L (j(\lambda^{(j)} - \zeta_p) \theta(\lambda^{(j)}, \zeta_p))}, \quad (20)$$

Where $L (L \leq N)$ is the number of statistical experiments where $\tilde{J}(\lambda) \leq \zeta_p$.

At the third stage, the obtained data is being processed, in particular, the definition of approximation polynomials with the use of the minor square method:

$$\lambda_i(p) = \alpha_i p^2 + \beta_i p + \gamma_i, i = \overline{1, c} \quad (21)$$

And the solution of the following tasks of single dimensional minimization:

$$\min\{\tilde{J}(\lambda(p)) | p \in [K - \Delta p, K + \Delta p]\}. \quad (22)$$

(22) The optimal solution of the task determines $\lambda_1^*, \lambda_2^*, \dots, \lambda_c^*$ (2) the optimal function of the control of the structure:

$$u_c^*(t) \approx u_c^*(\lambda) = \sum_{i=1}^c \lambda_i^* \omega_i, \quad (23)$$

By means of which the optimal trajectory $y^*(t)$ is calculated using the numerical solution system of the differential equations (12). On the basis of the results, the extreme value of the general criterion of optimality is $\tilde{J}^* = \tilde{J}(\lambda_1^*, \lambda_2^*, \dots, \lambda_c^*)$.

III. DETERMINATION OF OPTIMAL PARAMETERS OF TRANSISTOR AMPLIFIER

The task deals with and solves the dynamic problem for the determination of optimal parameters of the elements of transistor amplifier with common emitter according to the integral criterion of maximum approximation of output impulse with desirable impulse [1,7].

Let us analyze a one-cascade transistor amplifier with a common emitter and determine the optimal parameters of its elements, namely, resistors based on the criterion of maximum approximation of output impulse with desirable impulse.

$$J = \int_{t_0}^{t_k} [U_{out}(t) - U_{des}(t)]^2 dt \rightarrow \min \quad (24)$$

The electrical principal scheme of transistor amplifier is shown in fig.1.

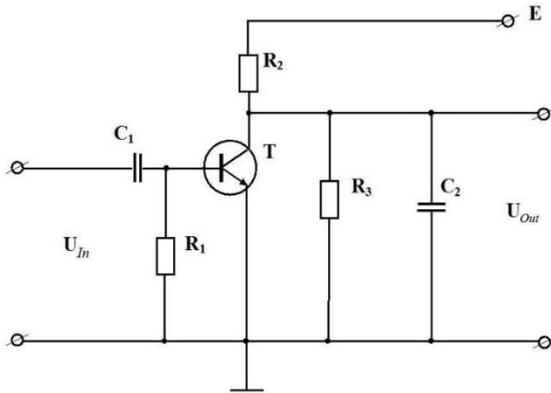


Fig.1

The optimization variable parameters are R_1, R_2, R_3 and fixed parameters are $C_1 = C_2 = 0,5 \cdot 10^5$ pF, $E=5$ V. In order to consider transistor in the amplifier's mathematical model the Ebbbers-Mall model [1] is used. The parameters of its equivalent scheme are:

- thermal current: $I_{TE} = 0,25 \cdot 10^{-12}$ mA,
 $I_{TC} = 0,6 \cdot 10^{-12}$ mA;
- charging capacity: $C_{ChE} = 0,15 \cdot 10^1$ pF,
 $C_{ChK} = 0,15 \cdot 10^1$ pF;
- temperature potentials: $m\varphi_{TE} = 2,60 \cdot 10^2$ V,
 $m\varphi_{TC} = 2,65 \cdot 10^2$ V;

- time constants: $\tau = 0,17$ nsec, $\tau_{imp} = 0,15 \cdot 10^2$ nsec;
- amplifying factor: $B = 0,5 \cdot 10^2$, $B_{imp} = 0,3$;
- leakage resistance: $R_{LkE} = R_{LkC} = 0,1 \cdot 10^{-6}$ k Ω ;
- base resistance: $R_B = 0,1$ k Ω ;
- collector resistance: $R_C = 0,02$ k Ω .

The optimal projection problem of the given scheme is as follows: it is necessary to select such nominal values of R_1, R_2 and R_3 resistors that at the time of the given U_{in} signal, the U_{out} signal should coincide with U_{des} signal to the maximum. (Fig.2). Thus, the optimization criterion is represented with the following functional:

$$J(R_1, R_2, R_3) = \int_{t_0}^{t_k} (U_{out} - U_{Des})^2 dt, \quad (25)$$

where t_0 is the moment of sending U_{in} signal, $t_0 = 10$ nsec, and t_k is the moment of the scheme reaction completion to the input signal, $t_k = 400$ nsec.

The parameters of U_{in} and U_{Des} signals are given and correspondingly equal:

$$t_{F1} = 15$$
 nsec, $T = 200$ nsec, $t_{F2} = 10$ nsec, $U_A = 2$ V;

$$U_0 = 0,2$$
 V, $U_M = 0,2$ V, $T_{imp} = 225$ nsec.

The degree of maximum approximation of output U_{out} signal with desirable U_{Des} signal is characterized in pic.2 by a hatched area the size of which is determined by a functional (2). Obviously, the minimization of the latter is the essence of the given problem.

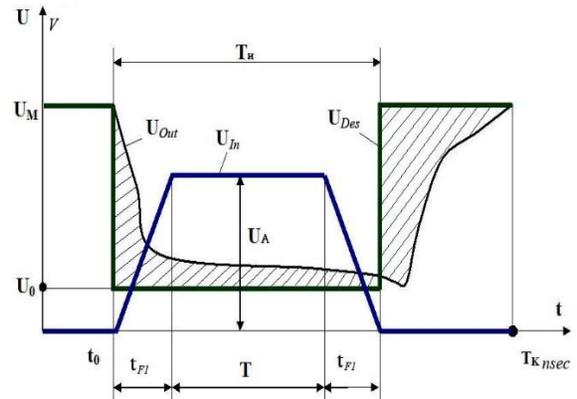


Fig. 2

The output U_{out} signal is determined by the result of numerical integration of the electrical scheme mathematical model ($\dot{y} = \varphi(y, \lambda, t)$, $y(t_0) = y_0$). As it is known, by means of state variable methods, the formation of mathematical model in computer is carried out in a planned way, in the automatic regime, and in the case of the transistor amplifier represented in pic.1 it can be expressed by the following system of equations:

$$\frac{dU_{out}}{dt} = \frac{1}{C_2} \left(\frac{1}{R_2} + \frac{1}{R_3} + \frac{1}{R_C} \right) U_{out} - \frac{1}{C_2 R_C} (U_C + U_E) + \frac{E}{C_2 R_2},$$

$$\begin{aligned} \frac{dU_{C1}}{dt} &= \frac{1}{C_1} \left(\frac{1}{R_1} - \frac{1}{R_B} \right) U_{C1} - \frac{1}{C_1 R_B} U_E + \\ &+ \frac{U_{In}}{C_1} \left(\frac{1}{R_1} + \frac{1}{R_B} \right), \\ \frac{dU_K}{dt} &= \left[C_{chK} + \frac{\tau_{imp} i_{TK}}{m\phi_{TK}} \exp \frac{U_K}{m\phi_{TK}} \right]^{-1} \cdot \\ &\cdot \left[\frac{1}{R_\sigma} (U_{In} - U_E - U_{C1}) - \frac{1}{R_C} (U_E - U_C - U_{Out}) - \right. \\ &- (B+1) i_{TE} \exp \frac{U_E}{m\phi_{TE}} + B_{imp} i_{TK} \exp \frac{U_K}{m\phi_{TK}} - \\ &\left. - \frac{1}{R_{LkE}} U_E + i_{TE} \right], \\ \frac{dU_E}{dt} &= \left[C_{chE} + \frac{\tau i_{TE}}{m\phi_{TE}} \exp \frac{U_E}{m\phi_{TE}} \right]^{-1} \cdot \\ &\cdot \left[\frac{1}{R_C} (U_E - U_C - U_{Out}) - (B_{imp} + 1) i_{TC} \exp \frac{U_K}{m\phi_{TC}} + \right. \\ &\left. + B i_{TE} \exp \frac{U_E}{m\phi_{TE}} - \frac{1}{R_{LkC}} U_C + U_{TC} \right] \quad (26) \end{aligned}$$

where U_{C1} is the voltage on $C1$ condenser, and U_C and U_E – respectively capacity voltages of collector and emitter passes of the transistor.

For the numeral integration of (3) system the Runge-Kutta method was used with the automatic selection of a step. At the same time, the calculation of the initial conditions in t point was accomplished according to the recommendations given in article [2, 8,].

The variability area of the R_1, R_2 and R_3 optimization parameters of the scheme was determined by the following inequality.

$$0.5(k\Omega) \leq R_i \leq 5.0(k\Omega), i = 1,2,3. \quad (27)$$

For the determination of the optimal parameters of transistor amplifiers with the central gravity method, the following program values have been selected:

- number of statistic experiments $N = 100$;
- number of Lebeg levels $K = 10$;
- compact factor $\rho = 20$;
- value of permissible error $\varepsilon = 0.01$.

The problem was solved on a modern PC and the following optimal solution was received:

$$\begin{aligned} R_1^* &= 4.1893k\Omega, \quad R_2^* = 0.9759k\Omega, \\ R_3^* &= 3.1011k\Omega; \quad J^* = J(R_1^*, R_2^*, R_3^*) = 33.4943. \end{aligned}$$

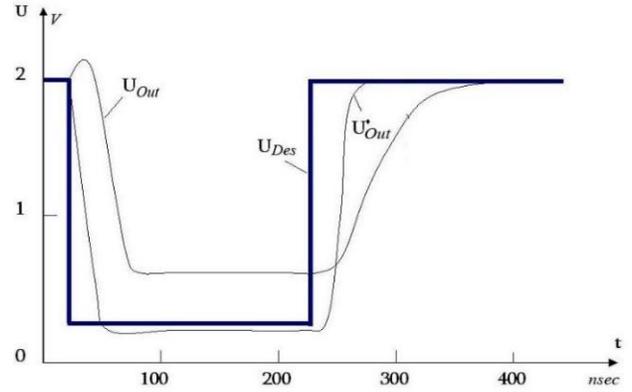


Fig.3

At the time of optimal values of the resistors, the degree of output $U_{Out}^*(t)$ signal approximation to desirable $U_{Des}(t)$ signal is graphically represented in fig.3. For the comparison, the same picture shows the diagram of output U_{Out} signal, which corresponds the following randomly selected values of transistors: $R_1 = 2.75k\Omega$, $R_2 = 4.81k\Omega$, $R_3 = 2.40k\Omega$.

CONCLUSION

Based on the gravity center methodology algorithm is designed for solving dynamic problems of parametric optimization for electrical circuits. Using the integral criterion to maximally approximate the output impulse to the desired impulse, the algorithm solves the problem of determining the optimal parameters for the common emitter transistor amplifier

REFERENCES

- [1] D. Lynn, C. Meier, D. Hamilton, Analysis and calculation of integrated circuits. v.1.2, M., Mir, 1969.
- [2] Kamien, M. and N.L. Schwarz. Dynamic optimization: the calculus of variations and optimal control in economics and management, Amsterdam: Elsevier Science, 1991; Dover edition, 2012.
- [3] M. Minoux, Mathematical programming: theory and algorithms, Wiley, 1986.
- [4] Goldfarb D. A Family of variable metric methods derived by variational means. Mathematics of Computation, 24, 1970.
- [5] J. Cea, Optimization - Theory and Algorithms. 1978, Bombay, India
- [6] Jibladze N., Imedadze T., Donadze M., Geometric Programming Tasks are solution by the center of gravity. Transactions. Georgian Technical University. Automated control systems, ISSN 1512-3979, № 2 (5) Tbilisi 2008, 37-41pp.
- [7] M. Donadze, Optimization criteria of electronic circuit design and algorithm of parameter definition, 2015 IEEE East-West Design & Test Symposium (EWDTS) Proceedings, Batumi, 2015, 396-400pp.
- [8] N.S. Bakhvalov, N.P. Zhidkov, G.M. Kobelkov, Numerical methods. M., Science, 1987.
- [9] Alekseev V.M., Tikhomirov V.M., Fomin SP. Optimal control. M. Science, 1979.

Intelligent Transport Systems as a Way to Improve the Quality of the Rail-Truck Multimodal Freight Transportation

Natalia Goncharova,
assistant at "Logistics and
Commercial Operations" Department,
Emperor Alexander I St. Petersburg State
Transport University,
St. Petersburg, Russia
nataliegoncharova@list.ru

Abstract—Presented research is dedicated to the problems of using automated systems in the area of rail-truck multimodal freight transportation. The share of multimodal transportation is steadily growing; the role of cooperation between enterprises of different modes of transport is increasing. At the same time, many problems arise in this area; the lack of a unified information environment for all participants in the transportation process is one of the most important of them. In our opinion, this problem could be solved by connecting road carriers and logistics intermediaries to the Computer and Information Center network of JSC "Russian Railways".

Keywords— *Computer and Information Center network of JSC "Russian Railways"; automated system for railway operations; automated system of operational transportation planning; multimodal freight collaboration.*

I. INTRODUCTION

Each mode of transport has its own advantages and disadvantages. In particular, when comparing rail and road transport, we find that competition between them is rarely justified. In most cases, mixed rail-truck transportation is more efficient. To improve this, it is necessary to correctly organize the collaboration of modes of transport, for example, on the basis of a unified transportation technology. The study of these problems is relevant for all countries in which railway transport is represented. This is especially applicable for Russia due to the large size of the territory. The common information base is a prerequisite for the development and implementation of such technology.

The purpose of this study is to analyze the interaction of information systems of rail and automobile transport. To achieve this goal, the research has the following objectives:

- the identification of problems arising in the organization of mixed rail-truck transportation;
- the analysis of opportunities to create a unified mixed rail-truck transportation technology;
- the research of foreign experience in the field of organization of mixed rail-truck transportation and the possibility of applying this in Russia;

- the study of the possibilities of Computer and Information Center network of the Russian Railways using for the mixed rail-road transportation technology development.

This problem is investigated in different countries; the general trend in these studies is the transition to the introduction of the "one window" principle in the multimodal transportation. A literature review [1-10] revealed that the most significant factors that impede collaboration are the lack of shared information, the lack of flexibility and compromise. Currently, there is a trend towards the creation of logistics centers to solve the cooperation problems.

II. THE RESEARCH IN THE AREA OF MULTIMODAL FREIGHT TRANSPORT COLLABORATION IN THE FOREIGN COUNTRIES

The rail-truck multimodal freight collaboration is now at an early stage of the development. The need to develop the rail-truck multimodal freight collaboration motivates the need for a comprehensive analysis of the drivers and the barriers collaboration of modes of transport. The most detailed analysis of these factors, in our opinion, is introduced in the research of Guo et al. [11].

Recently, the concept of multimodal door-to-door freight transportation has been developing in Western Europe [12]. The main participants in the international multimodal freight transportation are shippers, forwarders and logistic providers. The pre-haulage, main-haulage, and end-haulage components are the base for the freight transport network. On Fig. 1 we may see the enlarged scheme of this network developed by Mutlu et al. [13]. We divide the forms of collaboration into two types: vertical and horizontal. The vertical cooperation is organized between two objects that belong to different levels of the supply chain. The horizontal cooperation is possible in a situation where two or more non-affiliated organizations cooperate by sharing information. In this way, the objects are in the same level of the supply chain.

Guo and Peeta [14] consider it a difficult problem that enterprises of different modes of transport often use different information technologies that do not interact with each other. Information sharing is a necessary condition for multimodal freight collaboration [15].

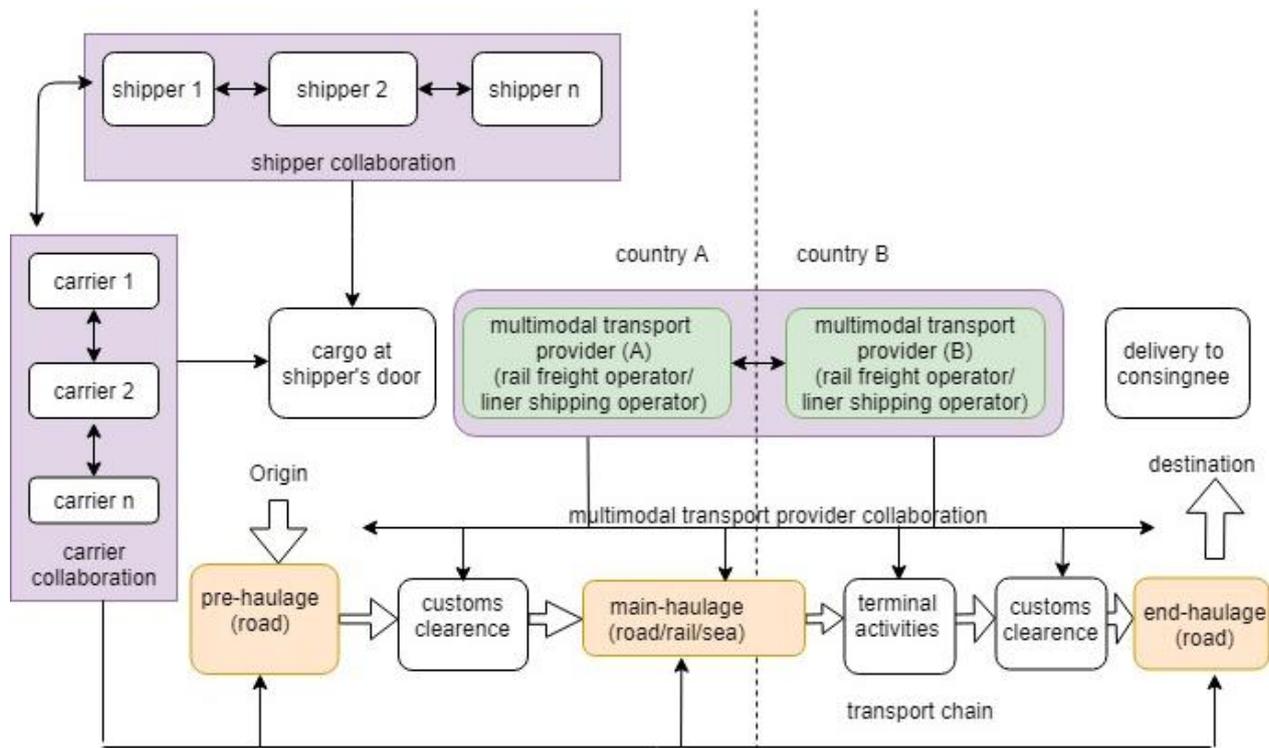


Fig. 1. Freight Transport Network.

For the development of multimodal freight collaboration, it is possible to use the Intelligent Transport Systems (ITS). ITS integrate different information sources (presented at Fig. 2). Accurate data is extremely important for a multimodal freight planning system.

The information for joint planning must be of high accuracy, because low quality data from one of several sources is enough to make the model results erroneous [16]. Hu et al. [17] propose to use the cloud-based decision support system for solving problems in the area of coordination and planning.

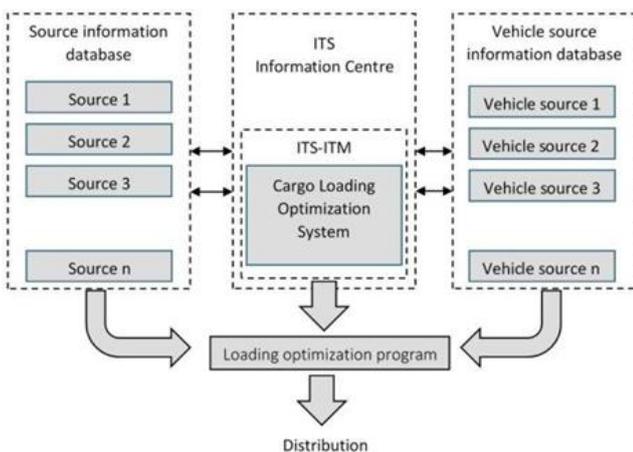


Fig. 2. ITS for cargo transportation optimization [18].

Heilig et al. [19] note that it is difficult to establish a centralized communication system, as transport providers compete with each other, and this makes the exchange of information between them unacceptable.

III. THE COMPUTER AND INFORMATION CENTER NETWORK IN FREIGHT AND COMMERCIAL OPERATIONS ON THE RUSSIAN RAILWAYS

A significant increase in the efficiency of freight and commercial operations on the railway transport was achieved through the automated control system creation. The automated system for railway operations allowed automating the collection, transmission and transformation of information, as well as the issuance of control actions. The automated control system operates on the basis of the Main Computing and Information Center, Computing and Information Centers for each railway, hub Computing and Information Centers, Computing and Information Centers for sorting and cargo stations, as well as other large enterprises of the Russian Railways system. The automated system for railway operations is a system consisting of a set of technical means of computing equipment, software, telecommunications and economic-mathematical methods, as well as the management apparatus, making decisions based on automated information processing. The automated system of operational transportation planning functions on all railways of the network, based on the Computer and Information Center network.

This system is intended for automated preparation and presentation of information about the transportation process to the heads and operational employees of railway departments, offices and stations for operational regulation of transportation process.

The machine model of transportation process on the railway polygon is the information basis for automated system of operational transportation planning for each railway. The information model reflects the current state of the operational work on the polygon. Basically, the system is used to service the station operational staff (operators of station's technology center and operators of commodity offices, station and shunting dispatchers), the stuff of railway departments (train and locomotive dispatchers, attendant on duty),

and also for operational and administrative departments of transportation services, heads of various levels.

At the first stage of automated system of operational transportation planning creation, models of trains, locomotives and special rolling stock were implemented. This system opened wide opportunities for improvement of operational work management on railways. It allows chiefs and operational staff to obtain a holistic view of the operational situation at controlled polygons at close to real time. The forecasting and operational planning of the forthcoming work became possible due to automated system of operational transportation planning. This system also made it possible to monitor compliance with technological discipline, to take operational measures to eliminate violations. The automated system of operational transportation planning ensured the issuance of a set of technological documents for each train to operational employees of stations and railway departments. It became the basis for the creation of new automated systems and complexes of tasks in the transportation process control system. The unification of the main design solutions in the field of information, software and technical support has opened wide opportunities for rapid replication and implementation of the system on the railway network.

System-wide means of the automated system of operational transportation planning were created centrally in the form of standard design solutions. This made it possible to unify the main processes of information handling in railway computing and information centers. The automated system of operational transportation planning implementation and creation ensured the construction of a reliable foundation for the computer network on the Russian Railways.

The next important point after the automated system of operational transportation planning introduction was the development of automated workplaces for employees of mass transport professions. The automated workplaces introduction has significantly reduced the time of operations of cargo and commercial work by solving the following tasks:

- reducing the time for employees to search for information required to complete documentation;
- minimizing the number of device failures;
- decreasing in the share of paper documents;
- reducing the number of errors that occur under the influence of the human factor.

The use of automated workplaces allowed to increase the productivity of operational staff in solving problems with a large number of accounting operations, increased the safety of train traffic and shunting operations and generally contributed to the improvement of working conditions. Then it became possible to build complex systems in which individual workplaces for the operational staff of the stations are interconnected and interact with each other, which increased the efficiency of cargo and commercial work. As a result of the addition the automated system of operational transportation planning by automated workplaces, the throughput and processing capacity of the stations was significantly increased due to the stable performance of tasks for loading, unloading, processing and passing the train flow, compliance with established standards and increasing the productivity of mass professions employees. At the same time it

was possible to achieve significant savings in operating costs.

Currently, JSC "Russian Railways" attempts to create the universal information service that provides planning and management of operational work on the basis of optimal interaction between all participants of the transportation process at all its stages. The largest developments in this direction are the third generation of automated system of operational transportation planning and the intelligent railway control system [20].

IV. CLUSTER COMPUTING CENTERS ON THE AUTOMOBILE TRANSPORT IN RUSSIA

The creation of automated control systems for the transport process on automobile transport began with the creation of cluster computing centers, in the continuation of this work, computer centers for collective use were created. The prerequisites for this were a high-quality build-up of the mathematical apparatus, the creation of software systems, and the emergence of opportunities for docking the tasks of planning the transportation process with the information-computing complex. The introduction of cluster computer centers into the activities of road transport enterprises has greatly simplified the solution of the following operational planning tasks:

- the determination of shortest distances between points of the transport network;
- the creation of transport routes;
- the optimal fixation of routes for trucking companies;
- the creation of tasks for drivers;
- the performance of cargo transportation on routes.

A significant drawback of the cluster computer centers was that it had very limited capacity to respond to the rapidly changing operational environment. In order to ensure the operational management of the transportation process, reliable information from the line received through the communication circuits was required.

The improving of the operational management in the transportation process in the 1980s was made possible due to the appearance of personal computers. On the basis of personal computers, automated workplaces for the staff associated with the transportation process management began to be created. The operational planning of road freight transport solved a finite number of problems arising from the use of various schemes for delivery of goods to the consumer. The use of cluster computer centers helped to solve such important ones as the determination of the shortest distances and the routing of field and small-batch shipments, the calculation of the minimum time for delivery of cargo.

The start of the use of automated workplaces in the activities of the cluster computer centers allowed obtaining the following results:

- minimize the downtime of rolling stock and cargo handling facilities arising from the inconsistency of their work;
- reduce the unevenness of freight and commercial work processes;
- reduce the number of errors that occur under the influence of the human factor;

- improve the quality of the creation of tasks for drivers and simplify the assessment of the degree of their implementation;

- improve the system for evaluating the efficiency of labor of automobile transport staff.

The fundamental changes in the economy of the country caused a breakdown in communications between the trucking companies, which made it impossible for the cluster computer centers to continue to function. In the absence of state centralized planning of enterprises, such centers have extremely limited possibilities of application.

V. PROSPECTS FOR USING THE COMPUTER AND INFORMATION CENTER NETWORK OF THE RUSSIAN RAILWAYS IN MULTIMODAL TRANSPORTATION

For the purpose of timely execution of accepted applications and unimpeded transfer of goods to other modes of transport (at the organization of multimodal transportation) JSC "Russian Railways" now carries out continuous planning of cargo transportation. The Computer and Information Center network has enabled the transition to a new progressive technology of continuous planning in railway transport.

At this stage, mixed rail-road transportations (not carried out on a uniform document) are massive. In the organization of this transportation process, started on the railway transport, continues after the transfer of cargo to the automobile transport. The automobile transport is the most flexible and mobile component of the transport system.

The integration of motor transport enterprises and the coordination of their activities with the Computer and Information Center network would allow to successfully solving the following problems of transport logistics:

- the ensuring technical and technological interconnection of participants in the transport process;

- the ensuring the technological unity of the transport and storage facilities;

- the joint planning of transport and warehouse processes.

The existing powerful Computer and Information Center network of JSC "Russian Railways" could become the basis for creating a computerized system for monitoring and planning multimodal transportation. The main results of this system will be the following:

- the reduction of cargo delivery time;

- the reduction of downtime for wagons and other vehicles;

- the acceleration of container turnover.

At the first stage, the Computer and Information Center network would help to solve a number of technological problems in the field of the organization of an integrated system for the operation of rail and automobile transport:

- the development of coordinated contact schedules for interacting modes of transport, consignors and consignees;

- the preparation of schedules of arrival and departure of different modes of transport, which are interrelated with the interests of shippers and consignees;

- the organization of complex technological processes in large hubs.

The creation of a unified information system of the transportation process operation will ensure the improvement of the quality of information support by forming a unified database of reliable data based on the consolidation of various information sources, facilities and events of the transportation process.

The creation of a universal information service would also increase the accuracy of the obtained data on the transportation process with the possibility of analyzing its history using archival data, which would simplify the process of identifying trends, dependencies and making forecasts.

VI. PROSPECTS FOR USING OF AUTOMATED CONTROL SYSTEMS OF LOGISTICS CENTERS

Currently, there is a tendency to create logistics centers to solve the problems of interaction between the participants of the transportation process. The problems of coordination and interaction between the participants of the cargo delivery chains can be solved by the use of automated control systems of logistics centers for the control and operation of cargo flow. There is a need for the organization of the transportation of goods, which provides for a uniform technological planning chain covering trunk transportation, coordination of the work of cargo yards, and management of local cargo delivery by the automobile transport.

The timeliness and the coherence of deliveries are ensured through operational interaction with shippers, consignees, owners of transport infrastructure, as well as transport companies. Through automated control systems of logistics centers, it is possible to track all stages of planning and execution of transportation, timely inform the participants of the supply chain about the situation in general, about the planned and actual time of arrival of the goods (presented at Fig. 3).

The information system itself receives data on the movement of goods by rail through exchange channels from the information systems of JSC "Russian Railways", in particular, electronic bill of lading and automated system of operational transportation planning. This will ensure the relevance of information, the possibility of timely submission of vehicles for the export of arriving goods; eliminate the overstock of freight yards.

To ensure the efficiency of management and control, all trucks are necessarily equipped with GLONASS navigation and communication equipment and connected to the automated control system of logistics centers.

The organization of operational information exchange between automobile and rail transport makes it possible to quickly respond to changes in traffic. Clear coordination of supply and operation of vehicles depends on the situation in the functioning of rail transport.

The problems of the coordination and interaction between the participants of the cargo delivery chains are often solved not through the use of automation methods and the creation of flow optimization algorithms, but through the personal interaction of logistics service consumers with leaders and managers of providers (often to the detriment of the interests of company-owners).

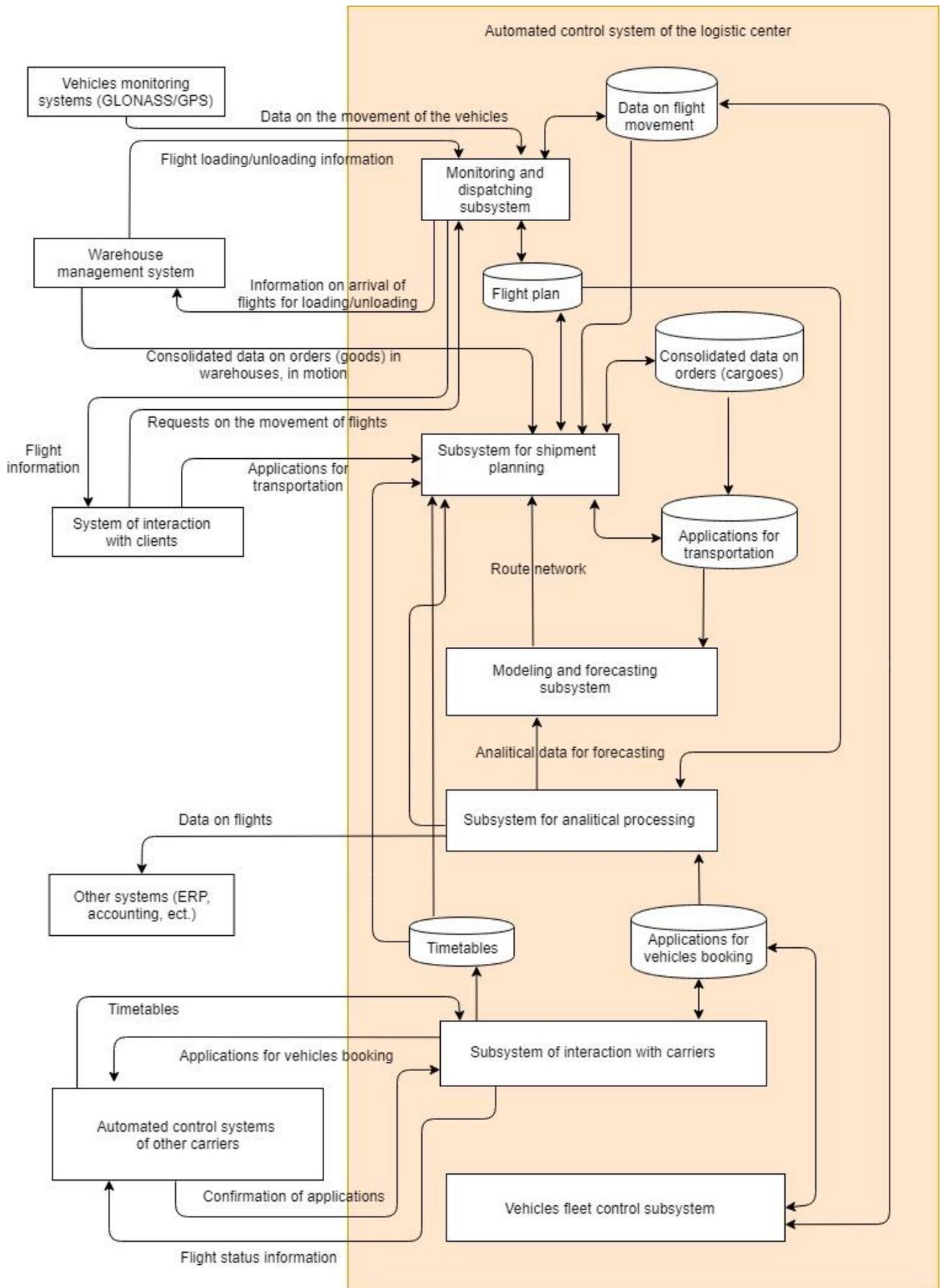


Fig. 3. The automated control system of the logistics center

In connection with this, the introduction of automated systems in terminal operations often encounters resistance of these categories of management personnel. The share of human participation in operation and exploitation processes should be reduced significantly, up to its attention to the role of an observer of the ongoing technological process [21, 22].

VII. CONCLUSION

In modern conditions, the share of multimodal transport is constantly increasing, and therefore the need for coordination of the various modes of transport is growing. The creation of a network of uniform centers managing the transportation process, based on the existing Computer and Information Center network of the Russian Railways, would dramatically improve the efficiency of logistic processes. At the moment, this process is difficult due to the lack of coordination of actions of various transport enterprises owned by different owners. Russian Railways could create a multimodal transport management system on the basis of the existing network, which would significantly improve their quality.

An urgent task at the next stage could be the development of a general algorithm for operational planning of the transportation process, based on a logistic approach. When forming such an algorithm, it is necessary to take into account the variety of options for interaction within the classical scheme (supplier – carrier – consignee) and, in the long term, the widespread use of more complex schemes of transport organization (with the inclusion of logistics intermediaries in the classical scheme).

The interaction between road and rail transport in multimodal transportation is now widespread. The share of multimodal transportation is steadily growing; the role of cooperation between enterprises of different modes of transport is increasing. At the same time, many problems arise in this area; the lack of a unified information environment for all participants in the transportation process is one of the most important of them. In our opinion, this problem could be solved by connecting road carriers and logistics intermediaries to the Computer and Information Center network of JSC "Russian Railways". The advantage of this way is compliance with the principle of logistics costs minimizing: there is no need to create new software products, new information networks, only to connect new participants to the existing network. On the basis of the Computer and Information Center network of JSC "Russian Railways", it is possible to organize the mutual exchange of information on a gratuitous basis, which would reduce the costs of all participants and promote the growth of cooperation. In the current situation in Russia, this way is the most effective for strengthening cooperation and solving the problems in the area of multimodal transportation.

REFERENCES

- [1] A. Cyril, Ö. Cagdas, J. Mandar, and N. Bernd "Analyzing the Potential of Future-Internet-Based Logistics Control Tower Solutions in Warehouses", Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI 2014), Qingdao, China, October 8-10, 2014, pp. 452-457, doi: 10.1109/SOLI.2014.6960767.
- [2] Y. Sun, and P. Schonfeld "Holding Decisions for Correlated Vehicle Arrivals at Intermodal Freight Transfer Terminals", Transportation Research, 2016, Part B, issue 90, pp. 218-240, doi: 10.1016/j.trb.2016.05.003.
- [3] M. R. Jabbarpour, H. Zarrabi, R. H. Khokhar, S. Shamshirband, and K.-K. R. Choo "Applications of Computational Intelligence in Vehicle Traffic Congestion Problem: a Survey", Soft Computing, 2018, vol. 22, issue 7, pp. 2299-2320, doi: 10.1007/s00500-017-2492-z.
- [4] N. Marković, Ž. Drobnjak, and P. Schonfeld "Dispatching Trucks for Drayage Operations", Transportation Research Part E: Logistics and Transportation Review, 2014, vol. 70(C), pp. 99-111, doi: 10.1016/j.tre.2014.06.016.
- [5] R. Šakalys, and N. Batarlienė "Research on Intermodal Terminal Interaction in International Transport Corridors", Procedia Engineering, 2017, vol. 187, pp. 281-288, doi: 10.1016/j.proeng.2017.04.376.
- [6] W. Bożejko, R. Grymin, S. Jagiełło, and J. Pempera "Robust Tabu Search Algorithm for Planning Rail-Truck Intermodal Freight Transport". In: K. Saeed, and W. Homenda (eds.) Computer Information Systems and Industrial Management, CISIM, 2016, vol. 9842, Springer, Cham, pp. 289-299, doi: 10.1007/978-3-319-45378-1_26.
- [7] I. Harris, Y. Wang, and H. Wang "ICT Inmultimodal Transport and Technological Trends: Unleashing Potential for the Future", Production Economics, 2015, vol. 159, pp. 88-103, doi: 10.1016/j.jippe.2014.09.005.
- [8] J. Udomwannakhet, P. Vajarodaya, S. Manicho, K. Kaewfak, J. B. Ruiz, and V. Ammarapala "A Review of Multimodal Transportation Optimization Model", Proceedings of the 5th International Conference on Business and Industrial Research (IC-BIR), Bangkok, Thailand, May 17-18, 2018, pp. 333-338, doi: 10.1109/ICBIR.2018.8391217.
- [9] M. Mniif, and S. Bouamama, "A Multi-Objective Formulation for Multimodal Transportation Network's Planning Problems", Proceedings of 2017 IEEE International Conference on Service Operations and Logistics and Informatics (SOLI 2017), Bari, Italy, September 18-20, 2017, pp. 144-149, doi: 10.1109/SOLI.2017.8120985.
- [10] D. Md., and Z. Islam, "Barriers to and Anablers for European Rail Freight Transport for Integrated Door-to-Door Logistics Service. Part 1: Barriers to Multimodal Rail Freight Transport", Transport Problems, 2014, vol. 3, pp. 43-56.
- [11] Y. Guo, S. Peeta, and F. Mannering "Rail-Truck Multimodal Freight Collaboration: a Statistical Analysis of Freight-Shipper Perspectives", Transportation Planning and Technology, 2016, vol. 39, issue 5, pp. 484-506, doi: 10.1080/03081060.2016.1174365.
- [12] A.L. Osório, L.M. Camarinha-Matos, and H. Afsarmanesh "Enterprise Collaboration Network for Transport and Logistics Services". In: L.M. Camarinha-Matos, and R.J. Scherer (eds.) "Collaborative Systems for Reindustrialization. PRO-VE 2015. IFIP Advances in Information and Communication Technology, 2015, vol. 463, Springer, Cham, pp. 265-276, doi: 10.1007/978-3-319-24141-8_24.
- [13] A. Mutlu, Y. Kayikci, and B. Çatay "Planning Multimodal Freight Transport Operations: a Literature Review", Proceedings of the 22nd International Symposium on Logistics (ISL 2017), Ljubljana, Slovenia, July 9-12, 2017, pp. 553-560.
- [14] Y. Guo, and S. Peeta "Rail-Truck Multimodal Freight Collaboration: Truck Freight Carrier Perspectives in the United States", Journal of Transportation Engineering, 2015, vol. 141, issue 11, pp.1-11, doi: 10.1057/s41278-016-0046-4.
- [15] M. Cao, and Q. Zhang "Supply Chain Collaboration: Impact on Collaborative Advantage and Firm Performance", Journal of Operations Management, 2011, vol. 29, issue 3, pp. 163-180, doi:10.1016/j.jom.2010.12.008.
- [16] J. D. Suárez-Moreno, J. Garcia-Castillo, A. M. Castañeda-Velasquez, and A. F. Cardenas-Hurtado "Making Horizontal Collaboration among Shippers Feasible through the Application of an ITS", Proceedings of the 2nd Latin American Conference on Intelligent Transportation Systems (ITS LATAM), Bogota, Colombia, March 19-20, 2019, pp. 1-6, doi: 10.1109/ITSLATAM.2019.8721342J.
- [17] Q. Hu, B. Wiegman, F. Corman, and G. Lodewijks, "Critical Literature Review into Planning of Inter-Terminal Transport: In Port Areas and the Hinterland", Journal of Advanced Transportation, vol. 2019, doi:10.1155/2019/9893615.
- [18] N. Leemekanond, and F. Akagi "Logistics Transportation System Based on ITS Technology", Proceedings of the 6th IEEE Conference on Awareness Science and Technology (ICAST 2014), Paris, France, October 29-31, 2014, pp. 67-72, doi:10.1109/ICAwST.2014.6981835.
- [19] L. Heilig, E. Lalla-Ruiz, and S. Voß "Port-IO: An Integrative Mobile Cloud Platform for Real-Time Inter-Terminal Truck Routing

- Optimization”, Flexible Services and Manufacturing Journal, 2017, vol. 29, issue 3-4, pp. 504-534, doi:10.1007/s10696-017-9280-z.
- [20] A. Pavlovsky, “Optimally integrate” (in Russ.), Control Board, 2015, № 4, pp. 32-34.
- [21] A.M. Romanchikov, V.A. Gross, D.V. Efanov, and A.Y. Vasilyev, “Digitalization of Railway Transport in Russia Russia” (in Russ.), Transport Of The Russian Federation, 2018, № 6 (79), pp. 10-13.
- [22] D.V. Efanov, and G.V. Osadchy “Paradigms for Building Control Systems on Railroad Transport: from the Systems of Electrical Interlocking of Points and Light Signals to Smart Grid Train Movements Controlling Systems”, Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS`2018), Kazan, Russia, September 14-17, 2018, pp. 213-220, doi:10.1109/EWDTS.2018.8524809.

Remote Administration of Information Systems Via E-mail

Zaza Davitadze
Computer sciences Department
Batumi Shota Rustaveli State
University
Batumi, Georgia
zazadavi@yahoo.com

Gregory Kakhiani
Computer sciences Department
Batumi Shota Rustaveli State
University
Batumi, Georgia
gkakhiani@gmail.com

George Beria
Computer sciences Department
Batumi Shota Rustaveli State
University
Batumi, Georgia
beria.giorgil@gmail.com

Abstract—The paper deals with a client-server model of information system management that provides a server (Information System) management using so called transparent client architecture. This means that there is no need to install any additional software from the client side. The article describes the system architecture and tools that were used to develop the program. Preferences received in case of using this solution. The presented approach allows to simplify the access to information systems. As a result, it will be extended remote access to new levels of information systems as well as "built-in" and may be in the IoT (Internet of Things) systems. In order to identify the requirements for remote control systems, the flagship solutions used in the free software market were studied and analyzed. Then service executes received command and sends confirmation E-mail to a sender if necessary. Paper also describes the principles of the system, functional model and used instruments.

Keywords - Remote Control, Email, Python, IoT.

I. INTRODUCTION

Administration of modern information systems is associated with large expenses because it needs well qualified engineers even for easy routine work, sometimes for whole day [1]. Even more, sometimes the work requires to connect remotely for consulting or some modifications, which may be cause of more spending [2]. Recently, on this background remote control software packages have been developed. They are Radmin[3], Ammy admin [4], Anydesk [5], Team Viewer [6] and many more. However, all these packages require fast internet connection between Server and Client computers. Besides, these computers must have decent recourses to work with last versions of this software packages and almost in every case some modifications of network configurations have to be done, like opening specific ports to make software operation. Opening ports could be a danger for network. Some hackers may sneak in. Usually IT personnel need to interact with system software and run a variety of services that are only allowed at system level. Often, the server part of remote access solutions can be based on the use of standard software, for example, various browsers. In this case, using a web browser as a tool to connect, it should always be turned on and running, which in turn will hamper the user. Furthermore, modern browsers use too much resources of a computer. For this and some other reasons after examination of similar software it has been decided to use well known programming language Python. It should be noted here that recently IoT -Internet of Things has

developed really fast, that is the main principle of connecting and maintaining various devices to the Internet, such as smart houses, smart things, etc. Recently, devices working on the IoT architecture are becoming more and more popular [7]. Therefore, manufacturers are increasingly putting Python interpreters on a number of devices.

Our way of solving the described problem satisfies the following preconditions:

- As it seems, implementation of such solutions in system has to be easy to use.
- By the systematic researches, we deduced, that mentioned systems is mainly used to solve simple problems, due to the fact, that in the situations when staff has to deal with more problematic and challenging complications, they prefer to stay on site, observe and supervise changes during troubleshooting. All of this suggests that sometimes simple cmd commands is sufficient to solve particular problems.
- Research says that all of the three solutions should have the logging system, which constantly, simultaneously saves the line of actions performed by program. With the mentioned way, finding the reason of program failure and resolving in the exact period of the time will be much easier than usual.
- The remarkable note is that, the main audience that uses such software systems have necessary skills to perform simple tasks on E-mail.
- The program should not require a large amount of resources from a computer or information system
- Multiplatform is also one of the remarkable factors.
- High bandwidth or network connection shall not be required for connection

Just as the functionality of any technical device depends on its design, it can also be said that this assertion is highly justified with the design of the software, especially the means of interacting with its user. It is obvious, that the ideal Progressive product must be absolutely transparent and while using that program, user must not be aware of its existence or working with software should be possible only with low-levelled qualified personal [8]. However, there is a tendency to slowly simplify the use of relevant tools in information sharing.

Transparent Interaction Methods Same as Transparent Interface is a very active area of research that involves handwriting and gesture recognition for understanding voice commands (manipulating electronic information using physical objects) and manipulating interfaces, interaction modes. [9]

Restoring of archived file function works with the same structure. Due to the fact that, downloading from external links requires both link verification and additional network load, here are some complications that may also be used for the realization of other network activities in the future (eg IoT systems management).

II. MAIN PART

Based on the above listed requirements, we considered that the most optimal solution is information system management through SMTP / POP3 [10] protocols. The concept described by us was developed using Python. Which has interfaces to all popular platforms [11]. It also has special libraries which will ensure its work with E-mail exchange services.

Service is installed on a server computer. It can read E-mails sent with SMTP/POP3 and download attachments or files from specified links. Service can also run Python file or shell commands.

This approach, except that it satisfies all of the above requirements allows us not to "attach" the administrative Staff to workplace and to use this method in a wide range of information systems.

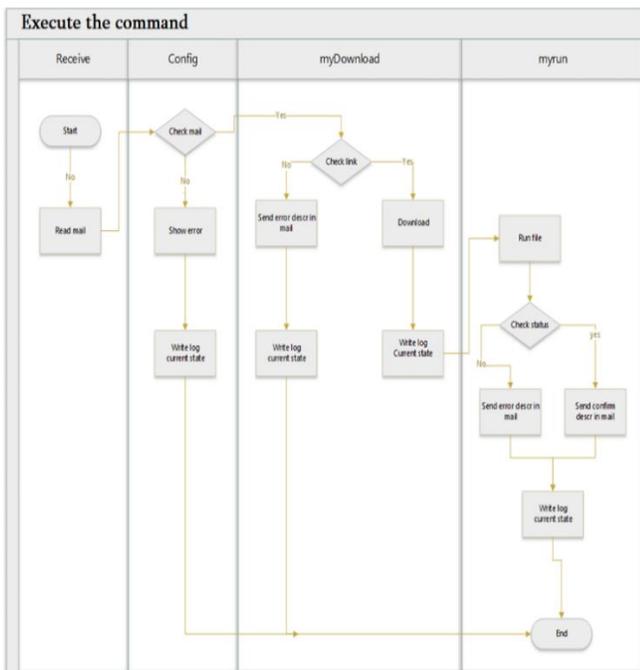


Fig.1. Functional scheme of system command execution

It is clear that each remote-control system involves receiving and processing tasks on the side of the managed system (in our case, this is the Server). In this case, tasks can be not only predefined commands embedded in the control system, but also as direct commands of the managed operating system. Therefore, the functionality of the solution can be extended

many times due to the deployment of software on the Server side.

The service running on the server is actually an e-mail client that makes a computer to run commands sent to specific E-mail address which is specified while installing service.

One of the commands is to download file and run it. The appropriate functional chart is shown in Fig. 1.

Before executing a command, service is constantly checking for new emails. Then it checks validity of email and if sender is in the list of allowed emails. This avoids dangerous tasks that must be performed from unreliable sources.

For security reasons every event taking place while program is running such as checking validity or download file is logged. Time sender and some other details are written in the standard log file and therefore are not anonymous.

In the case of an unreliable source, the system records this fact in the internal system log, ignores the received commands and returns to the e-mail waiting mode. If an E-mail is from reliable sources it downloads file from the link which is in the email message and runs that file. If the link is not valid sender receives email saying "link is not valid please check and resend it." after every step is done sender receives confirmation E-mail.

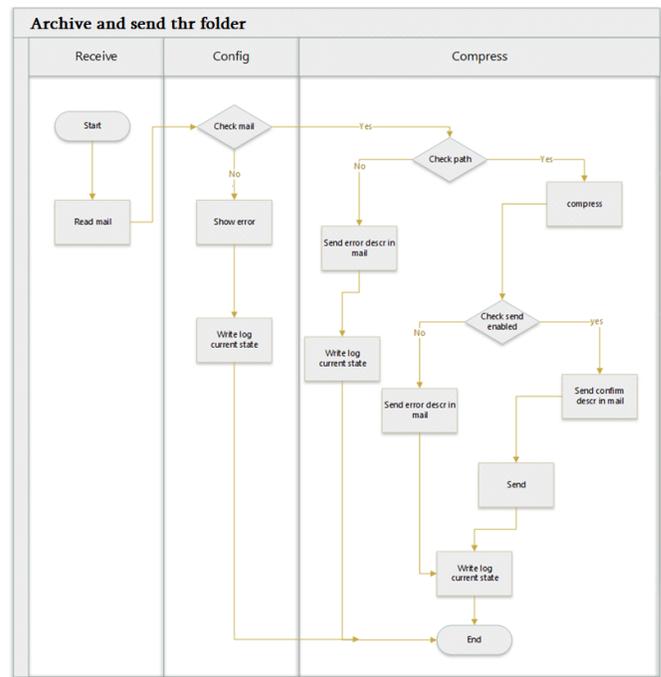


Fig.2. Functional scheme for archiving and sending command

Often some commands produce files which are requested to be sent to the sender of the task. For this reason, system has a functionality to compress files from specified path into .zip archive. Some email services such as Gmail or Yahoo don't allow unknown or dangerous files to be sent using their services. That is why files are compressed into zip archive. This action allows to send files inside a password protected archive (Fig. 2).

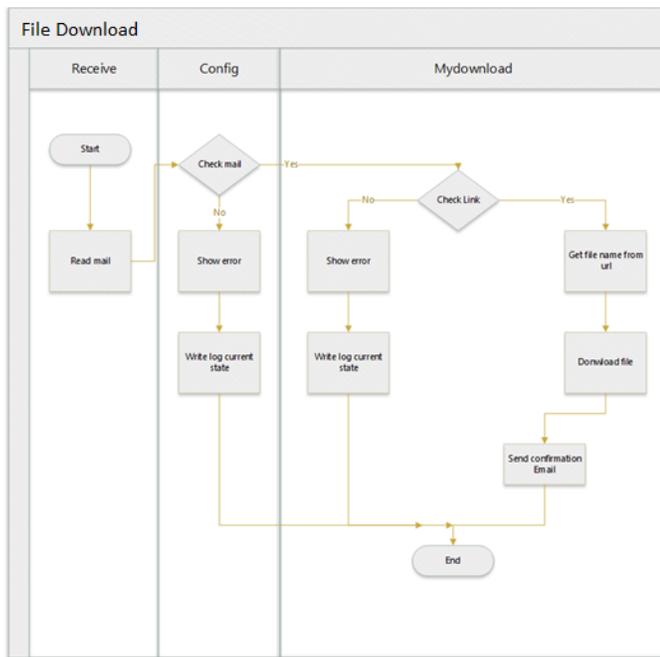


Fig.3. Functional scheme of download results

Sometimes files executed by the command are required to be sent to the user. It is advisable to archive these files. For this purpose we have added a new functionality to the system that archives the desired folder or files and sends them to the user specified address. Functional chart of the relevant process see Fig. 3.

This scheme similarly to the previous one, describes how service checks for validity of sender, writes every event in log file and only after validating the command is executed.

When E-mail is sent for predefined internal remote-control system commands syntax. At first it is needed to specify a function to which is referred by the writing its name. After that it is specified a parameter of function.

The syntax of the internal remote-control system commands used is as follows:

Command parameter

where the **Command** is an internal command of the remote-control system and *parameter* - text expression with parameters.

The following internal commands are implemented at this stage:

Screenshot DestEmail - allows to send Server computer screenshot on *DestEmail* E-mail address

Add NewEmail - register a new address in the remote-control system, *NewEmail* – new E-Mail address.

Execute Task - with administrator rights performs a *Task*, which is an external command (executable program or script)

Send File DestEmail - sends *File* to *DestEmail* – E-mail address

Download FileUrl - download files from *FileUrl*

Currently, this set of commands allows you to solve almost all remote-control tasks.

Service can execute external files like Python scripts or shell command files (.sh, .bat, etc.).

The created software package works under the object-oriented paradigm and consists of the following main classes:

1) *RCmail* - is responsible for the reading emails and execution of pre-prepared tasks in case they are requested. Some of the tasks are:

- compress (Archives the desired folder or (and) files)
- upload (upload the specified file to the predefined address)
- send file zipped (sends an archived file via email.)
- screen (takes screenshot and sends it to specified email)
- download (downloads the file from link and stores it at the specified path)

2) *RCMaillog* is responsible for logging processes in the service. It contains write and delete functions.

3) *PythonService* makes the program run in the background in the system as a service which is constantly running and reacts to external events. It also provides the program to connect with the operating system kernel programming interface.

Utilities which help to achieve windows service functionality are *win32serviceutil*, *win32service* Python packages. These packages are specifically designed to handle Windows services. They include functions (*StartService*, *StopService*, *RestartService*) that are required for Windows platform to recognize module as a service.

In Linux same functionality can be achieved easier. Here we can use *systemd* utility, which is System and Service Manager for Linux and Linux based systems. It will run as a first process and manage other processes. Using *systemd* we can monitor our services state.

Systemd is available for this distribution: Ubuntu, SUSE Linux Enterprise Server, Solus, Red Hat Enterprise Linux, openSUSE, Mint, Mageia, Fedora, Debian, Arch Linux, Parabola GNU/Linux-libre, Void Linux, lackware Knoppix, Gentoo Linux, Devuan, Android and some more.

To create the system service using *systemd*, we must specify unit file for this service. This file will be interpreted as a configuration file for our service. The file must be located in one of these paths:

- ~/.config/systemd/user/*
- \$XDG_RUNTIME_DIR/systemd/user.control/*
- \$XDG_RUNTIME_DIR/systemd/transient/*
- \$XDG_RUNTIME_DIR/systemd/generator.early/*
- ~/.config/systemd/user/*
- /etc/systemd/user/*
- \$XDG_RUNTIME_DIR/systemd/user/*
- /run/systemd/user/*
- \$XDG_RUNTIME_DIR/systemd/generator/*
- ~/.local/share/systemd/user/*

```
...
/usr/lib/systemd/user/*
$XDG_RUNTIME_DIR/systemd/generator.late/*
```

It's well known that the purpose of the interface design has long been to remove physical interfaces in the user's interaction with the human and computer. One of the important fields of this trend is "Automated capture" [12], which means remembering / saving recurring tasks so that it can be easily executed in the future.

While developing above described service following python packages were used. Those are: *E-mail, Os, Poplib, Smtplib, time, winsound, zipfile, datetime, HTMLParser, Requests, Winsound, win32service, win32serviceutil*. Used modules were checked for last standards compatibility.

Most importantly logs (Fig.4) are written for every event taking place while program is running. Their format is following:

YYY-MM-DD__HHMM, Task State, User, Task Name

Where YYY-MM-DD – date when function started, HHMM - time when function started, Task State – is the state of the execute (Begin or Finish), User – user who created task (similar valid E-mail), Task Name – internal function name, predefined external function name or executed commands.



Fig.4. Explanation of Logs.

For IoT systems one of the main part of architecture is IoT Cloud (Fig.5.). World almost every IoT devices are registered somewhere in the IoT [13] Cloud and run through third party servers.

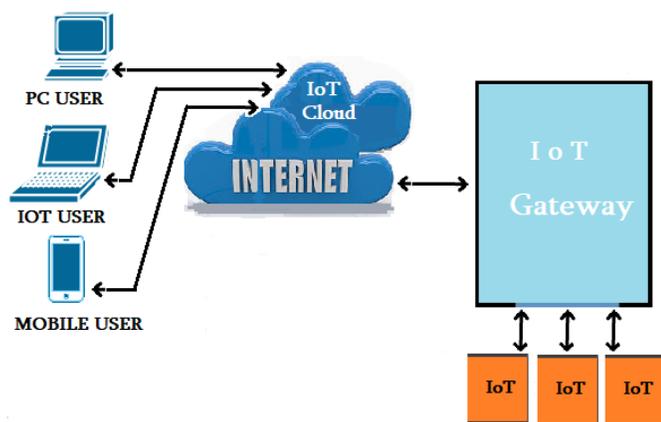


Fig.5. Normal architecture of IoT.

We have conducted a simple experiment in terms of remote control of items and to convey the concept below without of registration on the IoT Cloud and without of installation of any additional soft. We have used Arduino microcontroller with the

Led attached to it and sent an E-mail to turn on or turn off the Led. For turn on Led email text contain litter “y” (yes) and for turn off Led E-mail text contain litter “n” (no). Experimentally system reacted successfully. That means every time we sent a command Led was illuminating or disabled. In our case device can work independently from such servers. Also, if email client fails to send an E-mail from mobile phone, we can use pc browser to send the same E-mail.

General architecture of remote controlling of external devices via E-mail presented on Fig.6. IoT Cloud isn't here. Command for controlling external device from Computer to Device transmitted directly by internet without IoT Cloud.

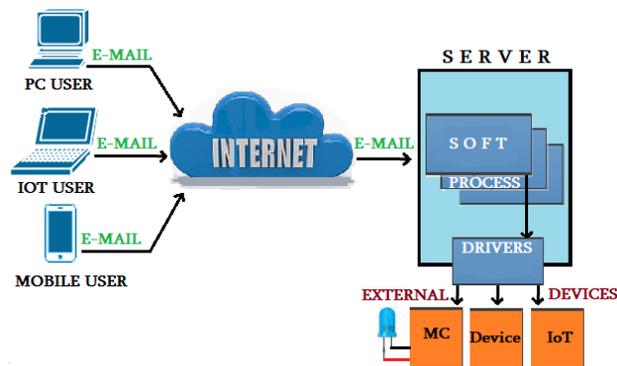


Fig.6. Aarchitecture of remote controlling of devices via email.

III. CONCLUSION

Presented solution allows provides access to administrative staff with transparent consumer interface in information systems as well as on desktops and embedded devices. Architecture allows the service to be used not only by humans but more automated programs can give orders too. The combination of built-in commands has a good perspective in order to become a specialized scripting language of administration.

REFERENCES

- [1] Nureni, Yekini. (2013). MANAGEMENT INFORMATION SYSTEM “Modern Perspective”.
- [2] Remote Access A Computer. <https://www.itarian.com/remote-access-a-computer.php>
- [3] Official Web Site of Radmin. <https://www.radmin.com/>
- [4] Official Web Site of Ammy Admin. <http://www.ammy.com/en/index.html>
- [5] Official Web Site of AnyDesk. <https://anydesk.com/en/>
- [6] Official Web Site of Team Viewer. <https://www.teamviewer.com/en/>
- [7] K Patel, Keyur & M Patel, Sunil & G Scholar, P & Salazar, Carlos. (2016). Internet of Things-IOT: Definition, Characteristics, Architecture, Enabling Technologies, Application & Future Challenges.
- [8] Gregory D. Abowd (1999) Transparent interaction, <https://www.cc.gatech.edu/fce/pubs/icse99/node9.html>
- [9] Jay Bolter, Diane Gromala (2014) Transparency and Reflectivity: Digital Art and the Aesthetics of Interface Design https://www.researchgate.net/publication/240360052_Transparency_and_Reflectivity_Digital_Art_and_the_Aesthetics_of_Interface_Design
- [10] Network Working Group (1996). Post Office Protocol - Version 3. <https://tools.ietf.org/pdf/rfc1939.pdf>

- [11] Dave Kuhlman (2013). A Python Book: Beginning Python, Advanced Python, and Python Exercises, https://www.davekuhlman.org/python_book_01.pdf
- [12] Beverly L. Harrison, Hiroshi Ishii, Kim J. Vicente, William A. S. Buxton (1995) Transparent Layered User Interfaces: An Evaluation of a Display Design to Enhance Focused and Divided Attention, <https://www.billbuxton.com/transparency.html>
- [13] Kang, Byungseok & Kim, Daecheon & Choo, Hyunseung. (2017). Internet of Everything: A Large-Scale Autonomic IoT Gateway. IEEE Transactions on Multi-Scale Computing Systems. 3. 206-214. 10.1109/TMSCS.2017.2705683.

Method of Indirect Steganographic Embedding on the Basis of Use of Functionality for Adaptive Position Number

Vladimir Barannik
Ivan Kozhedub Kharkiv National
University of Air Force,
KhNUofAF
Kharkiv, Ukraine
vbar.off@gmail.com

Dmitry Barannik
Kharkov National University of
Radio Electronics, KhNUofAF
Kharkiv, Ukraine
d.v.barannik@gmail.com

Nataliy Barannik
Ivan Kozhedub Kharkiv National
University of Air Force,
KhNUofAF
Kharkiv, Ukraine
Barannik_V_V@mail.ru

Abstract— In this article, a novel method for realization of hiding information based on special structural transformation of the positional number in the image is considered. Proposed method uses the structural dependencies between the elements of the image container.

Keywords—steganographic embedding, position number, information resources, information attacks, video files.

I. INTRODUCTION

Development and distribution of data telecommunication transmission media create potential treatments of information resources confidentiality violation. It is caused on the one hand by the increase of the importance of the transmitted data and on the other hand enhancement of means of the opponent for information attacks implementation.

The actual modern direction of ensuring confidentiality of information resources is use of computer steganography methods. In difference from cryptographic methods of information security, steganographic algorithms allow providing reserved imperceptible information transfer in the neutral container.

The most widespread methods are algorithms which use digital images as the container. Such distribution caused by the availability of extensive areas with psycho-visual redundancy in the image and wide dissemination of digital images and video files [1-7,10].

The existing methods of a steganography are realized by the direct and indirect embedding of information in the image container.

Methods of indirect embedding have advantages in comparison with direct embedding. But despite the advantages of the indirect approach, the existing algorithms don't satisfy up to the end requirements in case of information security. Such shortcomings are caused by using psycho-visual redundancy for embedding. In case of development of a method of steganography embedding to eliminate the existing defects, it is offered to use functional transformation for adaptive position number [8-9, 12-15].

II. MAIN PART

First, In the case of designing a steganographic transformation it is necessary to take into account the structural dependencies between the elements of the image container.

As a functional transformation that takes into account structural dependencies, it is proposed to use a function for an adaptive positional number, and as an element of an image container, a fragment of an image with dimensions of rows and columns [11,13].

Functional transformation for adaptive position number allows to reveal the structural regularities in the image caused by restriction for dynamic range:

$$\psi = \max_{1 \leq i \leq m} \{ c_{i,j} \} + 1, \quad j = \overline{1, n}.$$

Here $c_{i,j}$ - j -th element in i -th line of array F .

In the course of realization functional transformation for adaptive position number the fragment F of the initial image is considered as a set of the adaptive position numbers $\{C(j)\}$ consisting of elements

$$C(j) = \{ c_{1,j}; \dots; c_{i,j}; \dots; c_{m,j} \}.$$

Values of a code $K(j)$ will be defined as the sum of position number elements $C(j)$ on their weight coefficients $V_{i,j}$ on a formula:

$$K(j) = \sum_{i=1}^m c_{i,j} V_i.$$

Weight coefficients $V_{i,j}$ define by following formula:

$$V_i = \psi^{m-i}.$$

The second stage provides formation of a codegram $S(F)$ which includes an service component $S(\Psi)$ and information component.

Process of reconstruction of adaptive position number element $c_{i,j}$ on the basis of a code $K(j)$ is carried out on a formula

$$c'_{i,j} = [K(j)/V_i] - [(K(j)/(\psi V_i)) \psi]$$

In case of adaptive position coding, value of the reconstructed element $c_{i,j}$ of number $C(j)$ of a fragment F doesn't change in case of coding and decoding with various bases ψ and ψ' , i.e.

$$\begin{aligned} c'_{i,j} = c_{i,j} &= [K(j)/V_i] - [(K(j)/(\psi V_i)) \psi] \\ &= [K'(j)/V'_i] - [(K'(j)/(\psi' V'_i)) \psi'] = c'' \end{aligned}$$

Here $c'_{i,j}$ - the element of number $C(j)$ reconstructed on the basis of bases system Ψ ; $c''_{i,j}$ - the element of number $C(j)$ reconstructed on the basis of bases system Ψ' ; $K(j)$ - the code representation of number $C(j)$ created in basis of the bases Ψ ; $K'(j)$ - the code representation of number $C(j)$ created in basis of the bases Ψ' ; Ψ' - value of the modified element $c'_{i,j}$ basis.

It is offered to use property of unambiguity of decoding of adaptive position numbers during creation of a method of indirect steganographic embedding of special information [14-16].

Let estimate the amount of redundancy R that is inserted into the codogram as a result of indirect steganographic embedding. The value will be determined based on the following expression [17]:

$$R = q(S'(j)) - q(S(j))$$

or

$$\begin{aligned} R &= [\log_2 \cdot \psi^m] + 1 - [\log_2 \cdot \psi^m] + 1 = \\ &= [m \cdot \log_2 \cdot \psi'] + 1 - [m \cdot \log_2 \cdot \psi] + 1 = \\ &= [m \cdot \log_2 \cdot \psi + \log_2 k] + 1 - \\ &- [m \cdot \log_2 \cdot \psi] + 1 = \log_2 k. \end{aligned}$$

As follows from this expression, the value of redundancy R , which is introduced as a result of steganographic embedding, depends on the modification coefficient k . Then, in order to ensure that the minimum value is inserted during

the steganographic embedding process, the following condition must be:

$$(\log_2 k) \rightarrow 0.$$

Therefore, in order to reduce the level of introduced redundancy R , it is proposed to embed elements in binary representation $b_\xi \in [0; 1]$, and the modification coefficient is selected based on the following rule:

$$k = \begin{cases} 0, & b_\xi \rightarrow 0; \\ 1, & b_\xi \rightarrow 1. \end{cases}$$

Indirect embedding of an element b_ξ of the hidden message $B = \{b_1; \dots; b_\xi; \dots; b_v\}$ is offered to be carried out to the image container block by modification of the basis ψ_i on the basis of the following rule:

$$\psi' = \psi + k, \text{ where } k = b_\xi.$$

Here ψ' - the basis modified as a result of indirect steganographic embedding; k - modification coefficient.

At the following stage value of a code $K'(j)$ for number $C(j)$ with due regard for modified basis ψ' is calculated:

$$K'(j) = \sum_{i=1}^m c_{i,j} \psi'.$$

The third stage provides formation of codegram $S'(F)$ which includes an service component $S(\Psi')$ and information component $S'(j)$.

For ensuring additional resistance of the embedded data to the steganographic analysis of the malefactor it is offered to carry out preliminary handling of fragments before embedding [18]. This handling includes pseudorandom choice of the image-container fragments for chaotic embedding of bits of information sequence. Selection of fragments is performed on the basis of the chaotic sequence created by means of the following expression:

$$h_\alpha = 3,9 \cdot h_{\alpha-1} (1 - h_{\alpha-1}).$$

Here h_α - α -th an element of chaotic sequence H . Feature of such representation is need of the initial element h_0 value choice. Considering that value of an element h_0 can be calculated in the range $h_0 = \overline{0,0(0)1; 3,9(9)}$, key information will possess sufficient complexity for implementation of unauthorized matching [19,21,23]. Pseudorandom distribution of blocks in case of embedding can

demand the considerable computing resources caused by big definition of the image container. For ensuring decrease in computing complexity and reduction of the number of operations is offered to perform of preliminary handling chaotic distribution not of fragments, but their indexes (line items) in the image.

Such key information represents value of an initial element for creation of chaotic sequence. It allows defining pseudo randomly distributed blocks in a course of embedding [24]. This preliminary transformation includes pseudorandom selection of fragments for withdrawal of the embedded information by chaotic distribution indexes (positions).

Then process of withdrawal of embedded data will include the following stages (Fig. 1):

1. Extraction from a codegram information part (code) $K'(j)$ by means of the basis ψ' .
2. Restoration of initial number elements:

$$c'_{i,j} = [K'(j)/V'_i] - [K'(j)/(\psi'V'_i)] \psi'.$$

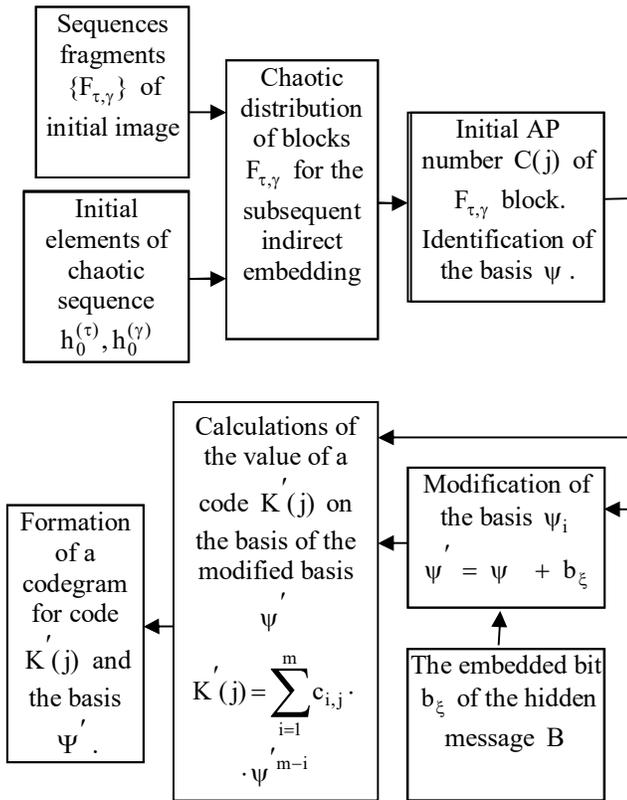


Fig. 1. The block diagram of indirect embedding

3. Identification of the initial basis ψ on a formula:

$$\psi'' = \max_{1 \leq i \leq m} \{ c'_{i,j} \} + 1$$

where ψ''_i - i -th basis of the restored basis Ψ'' .

4. Indirect withdrawal of the indirectly embedded bit b'_{ξ} . This stage is realized on the basis of comparison of the modified ψ' and restored ψ'' basis by the following expression:

$$b'_{\xi} = \begin{cases} 0, & \rightarrow \psi' - \psi'' = 0; \\ 1, & \rightarrow \psi' - \psi'' = 1. \end{cases}$$

or

$$b'_{\xi} = \psi' - \psi''.$$

Now we will consider the indirect steganographic transformation in case of not authorized access. In this case the malefactor has no key rule of embedding, and decoding will contain the following actions:

1. Extraction from a codegram information part (code) $K'(j)$ by means of the basis ψ' .
2. Restoration of initial number elements:

$$c''_{i,j} = [K'(j)/V'_i] - [K'(j)/(\psi'V'_i)] \psi'.$$

where $c''_{i,j}$ - i -th element of the reconstructed number $C''(j)$, as a component of the reconstructed fragment F at not authorized access.

Thus the developed method allows to carry out steganographic embedding and withdrawal of the built-in information bit.

III. ASSESSMENT OF PROPOSED METHOD

A comparative assessment of the amount of embedded information for the developed steganographic method and the existing steganographic methods of indirect steganographic embedding has been carried out. Among the existing methods of indirect steganography, the most common is the method of relative replacement of DCT values (Koch and Zhao method). This method is widely used due to the simplicity of its implementation and its wide applicability to hide data [21]. These advantages are due to the performance indicators of the method and the possibility of implementing indirect retrieval of embedded information without a prototype of the container image.

From the analysis of the research results of the developed method based on the program model, the following conclusions were formulated:

1) for the developed method, the amount of embedded information does not depend on the saturation of the image, but depends on the number of elements in the container;

2) the largest amount of embedded information for the developed method is observed in the case of the formation of

fragments for embedding by dimension 4×4 and takes the values:

for the highly saturated image - 9 KB;

for the slightly saturated image - 6.5 Kb;

3) for the developed method, the gain relative to the method of relative replacement of DCT values (Koch and Zhao) in the number of bits of embedded information ranges from 70% to 300%.

The results of the experiment for the visual assessment of reconstructed images with unauthorized access, in terms of the choice of bases of adaptive positional numbers from a fragment of the dimension 4×4 of the element are presented in the following images:

- highly saturated reconstructed image;
- slightly saturated reconstructed image.

From the visual assessment of the reconstructed images, the following conclusions were made:

- on the reconstructed images there are no visual distortions;
- the introduction of distortions does not depend on the saturation of the original container;
- the number of visual distortions does not depend on the dimension of the fragments formed by indirect steganographic embedding of service information.

An assessment of the sustainability of data embedded based on the developed method of indirect embedding in an attack using an active attack aimed at destroying the embedded message was made.

Evaluation of the developed method includes the following steps:

1. Perform direct discrete cosine transform.
2. Direct quantization with various factors of quality loss.
3. Inverse discrete cosine transform followed by rounding a real number.
4. Perform inverse quantization.
5. Comparative evaluation of the error bits of the embedded message that were correctly removed.

In the evaluation process, the values of code structures formed for images of various types are subject to attacks, namely:

- 1) highly saturated image;
- 2) low-saturated image.

The experiment is conducted under the following conditions:

- 1) embedding is carried out by modifying the bases of image fragments;
- 2) embedding is carried out for the three color components of the image under study;

3) the value of the quantization coefficient is chosen equal to $q=1; 2; 4; 10$.

Based on the results of the assessment, the following conclusions can be drawn:

1) for the developed method of indirect steganographic embedding, the number of error-free bits in the absence of attacks assumes a value of 100%;

2) for the developed method, the smallest amount of information correctly recovered 67.2% is observed for fragments with dimension 8×8 in the case of an attack of DCT with a quantization coefficient $q=10$;

3) the largest percentage of 70.4% by the number of error-free bits in the conditions of DKP attack and quantization $q=10$ with a step for the developed method is achieved when fragments are formed, by dimension 4×4 ;

4) in terms of the use of active attacks, the gain for the developed method with respect to the Koch and Zhao methods in terms of the number of correctly extracted data is on average from 10 to 25%.

IV. CONCLUSIONS

The method of indirect steganographic embedding of the built-in data on the basis of service data modification is developed. Indirect steganographic embedding is based on the following stages:

- identification of the basis for an image container fragment;
- embedding of bit of the hidden message by modification of the revealed basis;
- forming of code representation of adaptive position numbers of image fragment on the basis of the modified basis.

Scientific novelty. The method of indirect steganographic embedding of bit of the hidden message by modification of the basis of image container fragment of the is developed. In difference from other methods indirect embedding is performed by modification of the bases of image fragment elements with the subsequent forming code on the basis of the modified basis for adaptive position numbers.

The method of indirect steganographic withdrawal of the built-in bit of the hidden message on the basis of comparison of the initial and modified bases is developed. The mechanism of the return indirect steganographic provides:

- 1) recovery of initial fragment elements of the image container on the basis of modified bases system;
- 2) identification of initial bases system from a fragment of the image-container;
- 3) indirect withdrawal of bit of the hidden message by comparison of the initial and modified bases.

The method of the return indirect steganographic transformation on the basis of comparison of the initial and modified bases is developed. In difference from other systems, recovery of an initial fragment of the image container is performed for not authorized and authorized user in the presence of key information. It allows to build in bit of the hidden special information in image container fragment on the basis of service data modification.

REFERENCES

- [1] V.G. Gribunin, I.N. Okov, I.V. Turincev, Digital steganography [Cifrovaja steganografija], M.: Solon-Press, 2002, 272 p.
- [2] G.F. Konahovich, A.Ju. Puzyrenko, Komp'juternaja steganografija. Teorija i praktika, K.: MK-Press, 2006, 288 p.
- [3] S.V. Ablamejko, D.M.Lagunovskij, Obrabotka izobrazhenij: tehnologija, metody, primenenie, Minsk: Amalfeja, 2000, 303 p.
- [4] V. Barannik, S. Podlesny, A. Krasnorutskiy, A. Musienko, V. Himenko, "The ensuring the integrity of information streams under the cyberattacks action". East-West Design & Test Symposium (EWDTS), 2016. pp. 1-5. DOI: 10.1109/EWDTS.2016.7807752
- [5] A.V.Agranovski, A.V. Balakin, V.G Gribunin., Steganography, digital watermarking and stegoanalysis [Steganografiya, tsifrovyye vodyanyie znaki i stegoanaliz], M.: Vuzovskaya kniga, 2009, 220 p.
- [6] D. Bandyopadhyay, K. Dasgupta, J.K. Mandal, P. Dutta, "A novel secure image steganography method based on chaos theory in spatial domen", International Journal of security and trust management (IJSPTM) Vol 3, no 1, February 2014.
- [7] V. Barannik, A. Alimpiev, A. Bekirov, D. Barannik, N. Barannik, "Detections of sustainable areas for steganographic embedding". East-West Design & Test Symposium (EWDTS), (Novi Sad, Serbia, 29 sept. – 2 octob. 2017). 2017. pp 1-4. DOI: 10.1109/EWDTS.2017.8110028.
- [8] A.M. Al-Shatnawi, M. Atallah, "A new method in image steganography with improved image quality", Applied Mathematical Science, Vol. 6, no. 79, p. 3907-3915, 2012.
- [9] V. Barannik, A. Bekirov, A. Lekakh, D. Barannik, "A steganographic method based on the modification of regions of the image with different saturation". Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), 2018. pp 542-545. DOI: 10.1109/TCSET.2018.8336260.
- [10] O.I. Stasjuk, "Suchasni steganographichni metodi zahistu informacii" [Modern steganographic methods of information protection], Information protection, No. 1(50), pp. 56-63.
- [11] R.C. Gonzales, R.E. Woods, Digital image processing, Prentice Inc., New Jersey: Upper Saddle River, 2002, 779 p.
- [12] T. Hui, "An M-Sequence Based Steganography Model for Voice over IP", Communications, 2009. ICC'09. IEEE International Conference on 2009.
- [13] T. Husrev, Mahalinggam Pamkumar, Sencar Data Hiding Fundamentals And Applications. Content Security In Digital Multimedia, ELSEIVER science and technology books, 2004. 364 p.
- [14] M. Kutter, Digital Image Watermarking: Hiding Information in Images. PhD, Thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1999.
- [15] C. Lin Watermarking and digital Signature Techniques for Multimedia Authentication and Copyright Proytction, PhD Thesis, Columbia University, 2000.
- [16] L. Marvel, Image Steganography for Hidden Communication, PhD Thesis. University of Delavare, 1999. 115p.
- [17] M.B. Medeni, El.M Soudi, "A novel Steganographic Protocol from Error-correcting Codes", Journal of Information Hiding and Multimedia Signal Processing, pp.339-343, 2010.
- [18] V. Barannik, Yu. Ryabukha, V. Tverdokhlib, A. Dodukh, O. Suprun, D. Tarasenko, "Integration the non-equilibrium position encoding into the compression technology of the transformed images", East-West Design & Test Symposium (EWDTS), (Novi Sad, Serbia, 29 sept. – 2 octob. 2017). 2017. DOI: 10.1109/TCSET.2018.8336260.
- [19] Rybko B. Information-Theoretical Approach to Steganographic Systems / B. Rybko, D. Ryabko // *Proc. IEEE International Symposium on Information Theory, Nice, France, 2007. P.2461-2464.*
- [20] Z. Wang, A.C. Bovik, H.R. Sheikh, "Image quality assessment: From error visibility to structural similarity", IEEE Transaction on Image Processing, Vol. 13, pp. 309-312, 2004.
- [21] G.K. Wallace, "The JPEG Still Picture Compression Standard", Communication in ACM, vol. 34, no.4, pp.31 – 34, 1991.

Surface visualization of flexible elastic shells

Marina V. Byrdina

Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
byrdinamarina@mail.ru

Lema A. Bekmurzaev

Designing, technology and design
(of Affiliation)
Don State Technical University
(of Affiliation)
Rostov-on-Don, Russia
Bekmurzaev.l@yandex.ru

Mikhail F. Mitsik

Mathematics and applied informatics
Don State Technical University
Rostov-on-Don, Russia
m_mits@mail.ru

Svetlana V. Rubtsova

Mathematics and applied informatics
(of Affiliation)
Don State Technical University
(of Affiliation)
Rostov-on-Don, Russia
rubic-svetlana@yandex.ru

Abstract— The work is devoted to the description of the visualization of the volume-spatial form of a thin-walled shell structure, which is represented in the stress-strained state, which occurs when the shell is fixed along the upper edge and is freely positioned below the fastening boundary in the field of gravity and elasticity of materials. Without gravity, the shell is a straight circular truncated cone. The developed software module can be used in design and calculation of thin-walled shell structures for their non-linear deformation, as well as their visualization. The spatial shape visualization of the shell structure can be used to simulate various products, for example, conical antennas or products of the textile industry, flexible elastic shells in hydraulic engineering, etc.

Keywords— *Thin flexible elastic shell, visualization of shell structures, shell forming, variation problem, Maple and Embarcadero Red Studio packages*

I. INTRODUCTION

Thin elastic shells have the property of efficiency in terms of the consumption of materials for their manufacture. Shells are able to withstand heavy loads and insulate a technical object from an aggressive environment; they are easily flown around by a stream of air or liquid, and also have a relatively small mass. [1]. The lightness of the shells is one of the most important factors for many technical products and assemblies, and in many areas of technology and engineering is a vital requirement. Since thin elastic shells combine lightness with high strength, they are widely used not only in construction and light industry, but also in mechanical engineering, aircraft building and shipbuilding. [2]. Based on the functional purpose, the shell can be of the most varied forms and subjected to power and temperature effects, as well as to the action of solar radiation, the influence of water and air flows [3]. Modern computer geometry has many tools for building of complex volumetric spatial objects, analyzing their shape and design.

Existing works of famous scientific schools on determining the shape of cylindrical surfaces of S.N. Bulatov, V.A. Kozlov et al., solve the problem of forming the elastic shells loaded with external influences [4], or solve the problem of stability of shell structures [5, 6]. In this case, variation methods are also used, or the shape determination of the shell is realized with the help of solving a system of nine differential equations of the mechanics of a deformable solid body.

The development of computer-aided design of clothing is carried out in the school of V.D. Frolovsky and V.V. Landovsky. For modeling complex surfaces, variation methods based on finite difference schemes are also used [7, 8]. Currently, almost all the leading global companies in the field of software development for the fashion industry are engaged in equipping their computer-aided design systems with a clothing shaping module: Gerber (England) has the APDS-3D package, and PAD System (Canada) has the 3D Sample module. Firms Investronika (Spain) and Lectra (France) also declare about their developments.

The emergence of new composite materials requires the development of adequate practical methods for assessing physical-mechanical characteristics of shell structures and design schemes that are used in the description of stress strain state of the shell. To solve the problem of forming a thin elastic shell with a fixed upper edge is necessary to predict the elastic characteristics of the structure in the field of gravity.

The scientific meaning of the problem of determining the volume-spatial shape of the shell structure is that the elastic conical shell with a circumferentially fixed upper edge from the original shape (without taking into account gravity) due to the gravity is distributed over the shell making a free deformation. In this case, the gravitational forces in a certain position will be balanced by the elastic forces of the shell material. It is necessary to find such position of a construction.

The purpose of the work is to develop a method and a set of programs for visualizing the volume-spatial form of a thin-walled conical-type elastic shell in a stress-strain state, which occurs in the field of gravity and elasticity when the shell is fixed along the upper edge. Visualization of the shell form is implemented in the environments of packages of applied math programs Maple and Embarcadero Red Studio [9, 10].

Research tasks:

- 1) to set the task of determining the spatial shape of a thin-walled elastic shell of conical type with a fixed upper edge;
- 2) to show the solution of the boundary problem for the case of small deformations in the polar coordinate system;
- 3) to develop a software package in the Maple and Embarcadero Red Studio environments for the visualization

of the spatial shape of the elastic shell of large taper and for large deformations.

The novelty of the work: a numerical-analytical method is proposed to describe the shape of a thin elastic shell, which allows to determine the shell shape in the stress-strain state of high taper, obtain the number of folds on the shell, the geometry of the deflections and stress values depending on the geometric characteristics of the fabric, surface density and rigidity on the bend.

II. VISUALIZATION OF FLEXIBLE SHELL IN MAPLE

Visualization methods developed in software systems for commercial use are hardly accessible to the user for designing their own structures effectively. The study of forming thin flexible elastic inextensible shells with the help of modern mathematical packages of applied programs makes it possible to carry out quite complex calculations, but in practical application it is not always convenient for design engineers.

To calculate the shape of the prototype of a product, a 3D model of the structure under study is required; for this purpose, visualization programs in spaces of Maple π Embarcadero Red Studio for thin elastic shells having their own characteristics and errors in constructing surface elements and spatial curves are developed. The arising difficulties of describing the shape of surface elements are solved using computer geometry methods to graphically visualize the structural features of the surface of complex geometry, their docking and articulation.

To analyze the spatial shape of the shell structure in the stress-strain state, it is necessary to know the maximum crease folds (Fig. 1), maximum normal stresses, as well as the deflection and stress intensity at each point of the shell [11, 12].

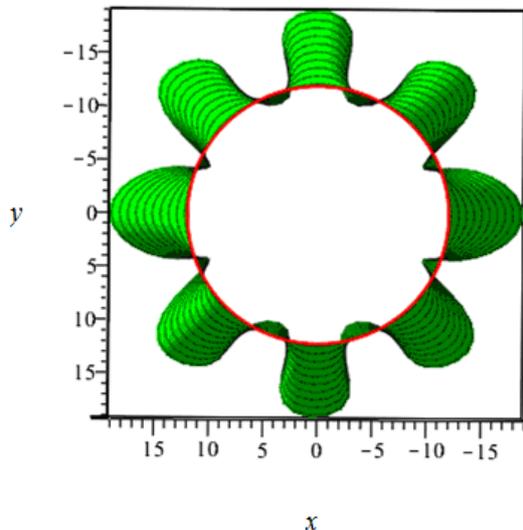


Fig. 1. Top view of a single-layer conic skirt of large taper

As one of the basic examples of shaping a thin elastic inextensible shell, consider the problem of determining the spatial shape of the surface of a single-layer conical skirt, which is conventionally made of a flexible elastic homogeneous material and its upper edge is fixed along the

perimeter of the waist. Figure 1 shows the visualization of a single-layer conical skirt of large taper (top view), obtained in the environment of Maple 2015 for the sizes of the product for a typical female figure.

The main dimensions of a typical female figure and the sizes of a conical skirt corresponding to it are presented in table 1.

TABLE I. ANTHROPOMETRIC DATA OF A TYPICAL FEMALE FIGURE AND VALUES OF CONE PARAMETERS

	<i>Dimensional signs</i>	<i>Value, cm</i>
1	Growth	164
2	Waist girth W_g	76
3	Projection distance from waist to hip h	20
4	Waist radius R_w	12.1
5	Hips Radius R_h	16.55
6	Skirt bottom radius R_{sb}	23.62
7	Skirt height H	51.73
8	Chest girth G_{ch}	96
9	Hip girth G_h	104
10	Skirt length L	53
11	The angle between the forming and the base of the cone β , radian	1.35
12	Generatrix length L	53
13	The length of the generatrix of the upper part of the cone R_v	55.62
14	Central sweep angle α , radian	1.37

Initially (without the influence of gravity), the skirt is a shell, which is a straight circular truncated cone for its middle surface (Fig. 2).

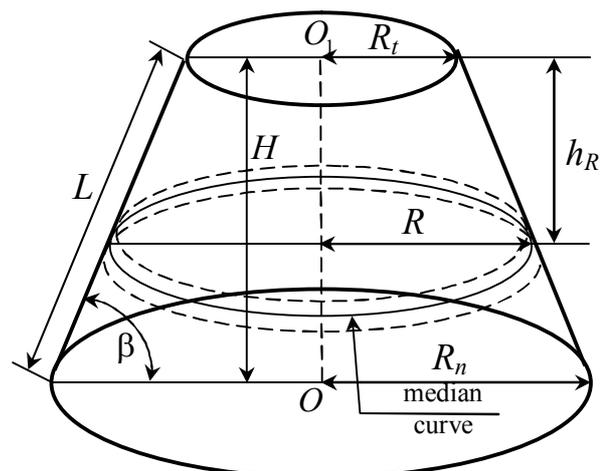


Fig. 2. Straight circular truncated cone with a splitting element

The spatial form of the shell (skirt) is determined by the conditions of its attachment along the upper horizontal border and free arrangement of the shell below the upper edge, the influence of gravity forces and the action of the elastic forces of the material from which the shell is made.

To build a mathematical model describing the shape of the product, we fix the upper edge of the shell on the device, which is a vertical tripod with a horizontal disc, the radius of which is equal to the radius of the waist (Table 1). Under the action of gravity and taking into account the elastic forces of the material from which the shell is made, its spatial form has the form shown in Fig. 3.

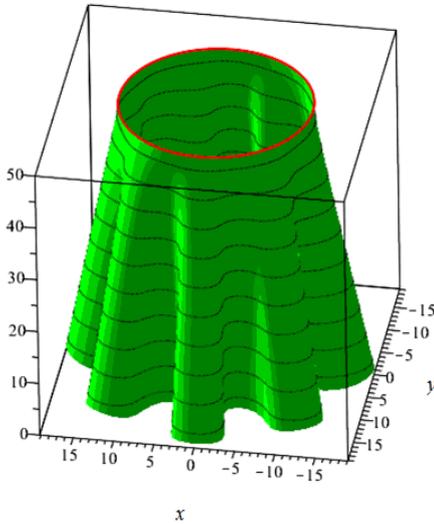


Fig. 3. Prototype of a flexible elastic shell with a fixed top edge

III. BOUNDARY VALUE PROBLEM TO DESCRIBE THE SHAPE OF THE SHELL

To find the shape of a thin elastic shell, we determine the potential energy enclosed in the shell if its upper edge is raised relative to the Oxy plane to a height L . Since the shell is raised relative to the horizontal surface, its original shape will be changed under the action of gravity. On the other hand, due to the presence of elastic forces of the material, the deformation of the shell will occur until the state of equilibrium of the forces of gravity and the elastic forces. The total potential energy concentrated in the shell should be minimal according to the principle of minimum potential energy [12]. This state of the studied structure can be described as

$$P_{grav} + \Pi_{ben} \rightarrow \min, \quad (1)$$

where P_{grav} – is the shell potential energy of gravity,

Π_{ben} – is the potential bending energy of the shell.

The task of determining the form of the shell can be solved as a variation method of finding the minimum of the potential energy of an element equal to the sum of the potential energies of gravity and elastic forces (1).

In the work [12] it was shown that for a shell of small taper, the potential energy of the position of a thin shell is determined by the formula

$$P_{grav} = \int \Delta P = \frac{g \cdot L}{\cos \beta} \cdot \int_0^{2\pi} \rho(\varphi) \cdot \left(L - \sqrt{L^2 - (r(\varphi) - R_t)^2} \right) d\varphi, \quad (2)$$

where g – is the acceleration of gravity;

$\rho(\varphi)$ – is the shell surface density;

$r(\varphi)$ – is the polar radius defining the distance of the shell section from the vertical axis.

The potential bending energy of a thin shell is determined by the formula

$$P_{ben} = \frac{L}{2I \cos \beta} \int_0^{2\pi} B(\varphi) \cdot (R_t - r(\varphi))^2 d\varphi. \quad (3)$$

where $B(\varphi)$ – is material stiffness;

I – is central moment of inertia.

Based on the principle of minimum potential energy and taking into account formulas (2), (3), we obtained the variation problem of minimizing the functional (4) with integral connection (5) [12].

$$\frac{g \cdot L}{\cos \beta} \int_0^{2\pi} \rho(\varphi) \cdot \left(L - \sqrt{L^2 - (r(\varphi) - R_t)^2} \right) d\varphi + \frac{L}{2I \cos \beta} \int_0^{2\pi} B(\varphi) (R_t - r(\varphi))^2 d\varphi \rightarrow \min. \quad (4)$$

$$\int_0^{2\pi} \sqrt{r^2(\varphi) + \left(\frac{dr}{d\varphi} \right)^2} d\varphi = 2\pi R. \quad (5)$$

The approximate solution of the variation problem is obtained in the form (6)

$$r(\varphi) = r_c + q \cdot \sin(\omega\varphi), \quad (7)$$

where r_c – is an average radius of a closed thin shell element;

q – is the amplitude of change of polar radius $r(\varphi)$;

ω – is oscillation frequency $r(\varphi)$.

The obtained numerical data to use the formula (7) is not enough. In this case, it is necessary to evaluate the shape of the shell and visualize its behavior. Such visualization of the deformation process allows the engineer to most clearly assess the stress-strain state of the structure [13, 14].

The shell visualization based on the obtained solution (7), its stress-strain state, is shown in Figures 1 and 3. In this case, the middle surface is the main surface describing the geometry of the shell; therefore, the scheme of its construction is of primary interest. A three-dimensional object is visualized by building its 3D model and then submitting it to the input of the graphics conveyors.

The results presented in the work were obtained on the basis of a software module developed in the space of the Maple 2015 package, intended for symbolic calculations. For visualization of shell structures in the developed software module, two coordinate systems are used:

(x, y, z) – is three-dimension Cartesian rectangular coordinate system describing the position of objects on the scene in which the shell model is built;

$(x(t), y(t), z(t))$ – is generalized cylindrical coordinate system in which deflections and information about the position of surfaces and spatial curves are specified [15].

When building a shell shape, it is assumed that the model is one color; all necessary attributes are transferred to the graphics processor once before rendering of the model. We believe that the light falling into a point is equally diffused in all directions. In other words, the illuminance is determined only by the density of light at a point on the surface, which depends linearly on the cosine of the angle of incidence.

When visualizing the calculation results, the main task is to define the transformation formulas for the generalized cylindrical coordinates of the shell into Cartesian rectangular, which take into account the shifts combining the selected center of reference with the beginning of the Cartesian coordinate system. When taking shifts into account, it is easier to implement the rotation of the shell model on the computer screen during its graphic visualization.

The Cartesian coordinates in Maple form a left-sided coordinate system and as they are arranged as follows (Fig. 3): the axis Ox is directed along the screen from right to left, the axis Oy is perpendicular to the screen and directed inwards from it, and the axis Oz is parallel to the screen and directed upwards. The generalized cylindrical coordinates are individual for each type of shell, only the z axis is always located in the vertical plane.

In Maple, it is possible to define a deflection surface, or a curve that returns the deflection value for a specific point in the selected coordinate system. A color model with a wide range of color options can be used to color display the deflection field and intensity of stresses on the surface of the shell.

Thus, a mathematical model of the shell deformation is considered, taking into account the nonlinearity of its shape, the boundary condition of fixing the shell along the upper boundary and the functional of the total deformation energy of the shell, which is the sum of the work of internal and external forces.

IV. CONCLUSION

Commercial software systems contain algorithms that are not known to the user, which creates difficulties for their effective use. There also may be errors in the construction of surface elements and spatial curves.

Presented in Fig. 1 and 3 visualization of the prototype shell structure in the stress-deformed state, obtained by using the developed software module, in a convenient form for designers, reflect the spatial-volumetric shape of the shell with the ability to select the field deflection and the number of folds.

The proposed program module can be used in the design and calculation of thin-walled shell structures of flexible elastic material, fixed on the upper edge and located freely in the field of gravity and elasticity. The module can also be used to analyze approximate solutions obtained by commercial software products for engineering calculations.

ACKNOWLEDGMENT

The research has been carried out at the expense of the Grant of the President of the Russian Federation for state support of young Russian scientists (MK-3403.2018.8).

REFERENCES

- [1] Peng Lan and Manlan Liu. Integration of Computer Aided Design and Analysis Using the Absolute Nodal Coordinate Formulation. 2011 Fourth International Conference on Intelligent Computation Technology and Automation. DOI: 10.1109/ICICTA.2011.48
- [2] A. L. Schwab and J. P. Meijaard. Comparison of Three-Dimensional Flexible Beam Elements for Dynamic Analysis: Classical Finite Element Formulation and Absolute Nodal Coordinate Formulation. *J. Comput. Nonlinear Dynam* 5(1), 2009, 10 p. doi:10.1115/1.4000320
- [3] A.M. Mikkola and M.K. Matikainen. Development of elastic forces for the large deformation plate element based on the absolute nodal coordinate formulation. *ASME J. Comput. Nonlinear Dyn.* 1(2), 103–108 (2006)
- [4] V.A. Kozlov. Theory and calculation of the conical shells of complex geometric structure: Dis. ... Dr. Phys.-Mat. Sciences: 01.02.04: Voronezh, 2003 245 p.
- [5] N. A. Alfutov, *Stability of Elastic Structures*, Springer-Verlag, Berlin, 2010. 335 p.
- [6] H. Ohmori, K. Yamamoto. Shape optimization of shell and spatial structure for specified stress distribution// *Memoirs of the School of Engineering. Nagoya Univ.*, 1998. Vol. 50. No 1. P. 1-32.
- [7] V.D. Frolovsky, V.V. Landovsky. Modeling of fabric based on particles method. *Proceeding of the International Forum isiCAD 2004. Novosibirsk. Ledas Ltd.* 2004. P. 224-229.
- [8] V.D. Frolovsky, V.V. Landovsky. Explicit and Implicit integration in the problem of modeling of fabric based on particles method. *Proceeding of 9th Korean-Russian International Symposium on Science and Technology. June 26-July 2, 2005. Novosibirsk State Technical University, Novosibirsk, Russia.* P. 596-600.
- [9] M.F. Mitsik, M.V. Byrdina, and L.A. Bekmurzaev. Modeling of developable surfaces of three-dimensional geometric objects. *Proceedings of 2017 IEEE East-West Design and Test Symposium, EWDTS 2017 2017. C. 8110086.*
- [10] Official site of Maple [Electronic resource]. – 2018. URL: <https://maplesoft.com/>
- [11] V.V. Novozhilov, K.F. Chernykh and E.I. Mikhailovsky. *Linear theory of thin shells. L. .: Polytechnic, 1991. 656 p.*
- [12] L.A. Bekmurzaev, M.V. Byrdina, E.V. Nazarenko. Research and modeling of thin shell formation // *Scientific and Technical Bulletin of the Volga region. 2014. № 4 p. 58-64.*
- [13] P.A. Zhilin. *Applied mechanics. Fundamentals of the theory of shells: Textbook.. St. Petersburg: Polytechnic Publishing House. University, 2006. 167 p.*
- [14] M.U. Nikabadze. The current state of multilayer shell structures // *Dep. in All-Russian Institute of Scientific and Technical Information RAS. 12.30.2002. No. 2289-B2002. 81*
- [15] Vasidzu K. *Variation methods in the theory of elasticity and plasticity / K. Vasidzu. - M. : Mir, 1987. - 542 p.*

??

AUTHOR INDEX

A

Abdullaev R. B. 157
Abdullayev V. H. 492
Abdulrahman Kataeba
Batiaa 513
Abramov O. A. 309
Adamov A. 502
Akishin B. A. 384
Aktemur T. Baris 43
Alevetdinova J. V. 350
Aleynikova O. 556
Amir Rikhtegar Ghiasi 78
Amirkhanyan K. 371
Andreeva V. 406
Antoshchuk S. 131
Anyutin N. 281
Aripov N. 531
Artyushenko V. M. 223, 243, 256
Avdalyan Narek 196
Aysu Aydin 43
Azarov A. 540

B

Badamchizade Mohammadali
552
Barannik D. 577
Barannik N. 577
Barannik V. 577
Baratov D. 531
Barch D. V. 484
Basharkhah K. 7, 23, 38
Behinfaraz Reza 552
Bekmurzaev L. 548, 556
Bekmurzaev L.A. 582
Belyi A. 201, 213, 544
Belyi A. A. 484
Beria George 572
Beyzanur Bora 119
Beyzanur Toprak 119
Biryukov Vadim N. 219, 252
Boldyrikin N. V. 400
Boutobza S. 13
Bugakova Anna V. 270, 301
Burdonov I. 445
Bureneva O.I. 148, 305
Butyrlagin Nikolay V. 263, 301
Byrdina M. 548, 556
Byrdina M. V. 582

C

Carlak Hamza Feza 450
Carlsson A. 502
Chastikov A. 466, 478
Chehraghi Seyedshehab 507
Cherckesova L. V. 376, 384,
392, 400, 410
Chernyshov S. 420
Chesnokov N. I. 410

Chibisov P. 274
Chumachenko S. 492
Costa A. 13

D
Davitadze Zaza 572
De Rossi Giacomo 507
Demakov A. 445
Denisenko Darya Yu. 263
Diachenko D. 207
Diachenko Y. 207
Djigan V. 29
Donadze M. 560
Donskoy D. Yu. 457
Drozd J. 131
Drozd O. 53, 131, 517
Drozdov D. G. 474
Drozdova I. I. 376
Dukanov P. A. 474
Dvornikov Oleg V. 270, 356
Dyka Z. 88, 97, 320
Devadze D. 48

E

Efanov D. 125, 136, 181, 484
Efanov Dmitrii V. 157, 162,
176, 289, 315, 343
Eghbali Zahra 68
Egorov S. 239, 326
Elaheh Mohammadi asl
Khasraghi 78
Elkayam M. 366
Evtushenko L. 461

F

Fazilov Sh.Kh. 191
Filippenko I. 488 Filippochkina
Anna O. 343

G

Galkin Yaroslav D. 270
Ghaemi Sehraneh 552 Gharibi
Wajeb 48
Ghasemy Seyyede Maryam 63
Ghavifekr Amir Aminzadeh 507
Goncharova N. 565 Gorbachov
V. 513
Gordon M. A. 544
Gordon Michael A. 309
Gören Sezer 43
Gourary Mark M. 153, 331
Grevtsev N. 274
Grigoryan H. 34
Grigoryan M. 34
Grimm C. 38
Grushin A. I. 474
Gül Nihal Güğül 119
Gulin A. I. 148, 305
Gulyaev A. A. 388

H

Hahanov Ivan 48
Hahanov V. 48, 108, 492
Hahanova Anastasia 492
Hahanova I. 108
Hakobyan Hovhannes H. 59,
103
Hoha M. 488

I

Ignashin Andrei A. 356
Ipek Seckin 43
Iskender Deniz 43
Ivanova Mariia V. 343
Ivanova O. 53

J

Jenihhin M. 23

K

Kabin I. 88, 97, 320
Kakhiani Gregory 572
Kaplanyan Taron K. 103
Karapetov E. 213
Karavay M. 108
Katunin Yuri V. 115
Kayar Ergin 450
Kelekhsaev D. 548
Khakhanova H. 108
Khanmohammadi Sohrab 552
Khilko D. 92
Khóroshev Valerii V. 162, 289
Khryashchev V. 470, 497
Klann D. 97, 320
Kluchka E. P. 457
Klyokta S. A. 410
Kokhanenko V. 556
Kokurin Joseph M. 176
Kondratenko A. 548
Kossachev A. 455
Kostanyan Hacob 34
Kostanyan Harutyun 34
Kostanyan Harutyun T. 59
Kotkova O. 513
Kovkin Alexey N. 309
Kraemer Rolf 430
Krstic Milos 430
Kulac Selman 337
Kulak E. 488
Kuliev Elmar V. 144
Kuperman A. 366
Kureichik V. M. 248
Kureichik Vladimir VI. 144
Kurganov V. 29
Kursityis I. 360
Kursityis Ilona O. 144
Kuzmin E. 424
Kuznietsov M. 131

L

Langendoerfer P. 97, 320
Langendörfer P. 88
Larionov R. 470, 497
Latypov R. 235
Lesnikov V. 466, 478
Letavin Denis A. 260, 267
Levchenko N. 73, 84, 395, 424
Liashov Maxim V. 356
Lighvan Mina Zolfy 68
Linkov V. 213
Litvinov M. A. 380, 388
Litvinova E. 48, 108
Lobodenko A.G. 376, 392, 400, 410
Logunova J.A. 248
Lukyanov A.D. 457

M

Malakhov M. 488
Malay I. 281
Malyshev A. 281
Man Ka Lok 492
Manaenkova O. N. 384
Margaryan H.34
Markevich A. V. 436
Martirosyan M. K. 523
Martynyuk D. 517
Martynyuk O. 517
Matrosova A. 406, 416, 420
Maunero N. 1
Melikyan V. 34
Melikyan V.Sh. 103, 523
Meskhidze Z. 560
Metelyov A. 478
Miroshnyk M. 488
Mishchenko A. 492
Mitsik M. 548, 556
Mitsik M.F. 582, 301
Mkhitaryan A. 34
Mohaghegh A. 7
Momjyan Arsen M. 103
Mona Saber Gharamaleki 78
Morozov S. A. 376, 384
Moskovskiy M. N. 380, 388
Movshin Anton A. 309
Musayelyan R. 34
Myachin V. 201
Myzdrikov N. Ye. 392

N

Nadolenko V. 229
Natskevich A. 360
Naumovich T. 466, 478
Navabi Z. 7, 23, 38, 63
Nikolaeva E. 416
Nosrati N. 23, 38
Okunev A. 73, 84, 395, 424
Orlov G. 92
Osadchy G. 125, 136, 181
Osadchy G. V. 484

Osadchy German V. 162
Ostanin S. 420
Ostrovskaya A. 470, 497

P

Pakhomov I. V. 380, 388
Palekha E. V. 384
Pankovsky B. E. 285
Pavlov V. 497
Petković Marko 457
Petrosyan Kamo 196
Pilipenko Alexandr M. 219, 252
Pilipenko I.A. 410
Pivovarov D. 125, 136
Poleskiy S. N. 285
Ponomarenko O. 513
Popa S. 13
Porksheyan V. M. 376, 384, 392, 400, 410
Prinetto P. 1
Prokopenko N.N. 148, 305, 252, 263, 270, 301, 356, 474
Provkin V. 416
Prozorov D. 527

R

Raik J. 23
Rajabalipanah M. 63
Rakhlis D. 488
Razumov P. V. 400
Reshetnikova I. V. 392
Revyakina Y.A. 400
Roascio G. 1
Rogdestvenski Y. 207
Rubtsova S. V. 582
Rumyantsev K.E. 171
Rusakov Sergey G. 153, 331
Ruziev D. 531

S

Sachenko A. 131
Sadeghi R. 23, 38
Safaryan O. A. 376, 384, 392, 400, 410
Safyannikov N. M. 305
Saglam Gurol 43
Sakharov I. A. 376
Samira Ahmadi Farsani 7
Sapozhnikov V. 125, 136, 181
Sapozhnikov Valerii V. 157
Sapozhnikov Vladimir V. 157
Sarmadi S. 63
Sassonker I. 366
Savchenko E. M. 474
Sedykh D. 201, 213
Sedykh D. V. 544
Sedykh Dmitry V. 309
Semenov A. 470
Semeonov I.Ye. 392
Serebryakov Alexander I. 219
Sergienko V. 488
Shakir H.H.Sh. 171

Shaporin R. 53
Shestovitskii D. 162, 201, 213
Shikunov Y. 92, 207
Shirokov I. 540
Shirokova E. 540
Sidorenko V. G. 436
Smirnov I. A. 400
Smirnov I. G. 380
Soofian Arezoo Beheshti 68
Stempkovskiy A. 229
Stenin Vladimir Ya. 115
Stepchenkov Y. 92, 207
Stepova H. 517
Stolov E. 235
Sugak L. 517
Sukhinets Zh. A. 148, 305
Sulima Y. 53
Surmacz T. 502
Svistunov G. 239, 326

T

Tatarinova A. 527
Tchekhovski Vladimir A. 270
Telpukhov D. 229
Teng Teng 181
Titov Alexey E. 270, 301
Trubchik I. S. 384
Tsyrunnikova E. 360
Tvardovskii A. 461
Tychinskiy V. 406

U

Ugurdag H. Fatih 43
Ulyanov Sergey L. 153
Ulyanov Sergey L. 331
Usatyuk V. 239, 326

V

Valiamova O. O. 148, 305
Vardanian V. 371
Vardumyan Arman V. 59
Varici Abdullah 43
Vasilenko M. N. 544
Vater Frank 97
Veleski Mitko 430
Vinarskii E. 461
Volovach V. I. 223, 243, 256
Yevtushenko N. 455, 461
Yildiz Abdullah 43
Yousefzadeh S. 23
Yukhnov V. I. 392
Yusupov O.R. 191

Z

Zahra Sattarzadeh Kalajahi 78
Zashcholkin K. 53
Zharov Michael M. 153, 331
Zhilin Victor V. 376
Zhuk Alexey A. 356
Zhukov A. I. 410
Zivkovic C. 38
Zmejev D. 73, 84, 395
Zuyev D.V. 544

Camera-ready was prepared by Chumachenko S.
Approved for publication: 07.09.2019. Format 60x841/8.
Relative printer's sheets: . Circulation: 100 copies.
Published by SPD FL Stepanov V.V.
Ukraine, 61168, Kharkov, Ak. Pavlova st., 311