ISBN: 978-1-7281-9898-9

2020 IEEE East-West Design & Test Symposium (EWDTS) Proceedings

Varna, Bulgaria, September 4 – 7, 2020

Proceedings of 2020 IEEE East-West Design & Test Symposium (EWDTS)

IEEE Catalog Numbers:
XPLORE COMPLIANT: CFP20DTW-ART
ISBN: 978-1-7281-9899-6

USB: CFP20DTW-USB
ISBN: 978-1-7281-9898-9

Additional copies of this publication are available from:
Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com

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

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 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

IEEE EAST-WEST DESIGN & TEST SYMPOSIUM 2020
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

18th IEEE EAST-WEST DESIGN & TEST SYMPOSIUM (EWDTS 2020)
Varna, Bulgaria, September 4-7, 2020

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 2020 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 Safely
- 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 Varna – is the largest city in northeastern Bulgaria, located along the Black Sea coast and Varna Lake. Varna is the administrative center of the municipality and the region, an attractive international educational center. The city has a rich cultural and historical heritage.

Because of the COVID-19 pandemic the IEEE EWDTS 2020 will be held online.

CONTENTS

A Review of Particle Detectors for Space-Borne Self-Adaptive Fault-Tolerant Systems

Marko Andjelkovic[1)], Junchao Chen[1)], Aleksandar Simevski[1)], Zoran Stamenkovic[1)], Milos Krstic[1), 2)], Rolf Kraemer[1), 3)]
[1)] *IHP – Leibniz-Institut für innovative Mikroelektronik, Frankfurt Oder, Germany*
[2)] *University of Potsdam, Potsdam, Germany*
[3)] *Brandenburg University of Technology, Cottbus, Germany*
{andjelkovic, chen, simevski, stamenko, krstic, kraemer}@ihp-microelectronics.com

***Abstract*—The soft error rate (SER) of integrated circuits (ICs) operating in space environment may vary by several orders of magnitude due to the variable intensity of radiation exposure. To ensure the radiation hardness without compromising the system performance, it is necessary to implement the dynamic hardening mechanisms, which can be activated under the critical radiation exposure. Such operating scenario requires the real-time detection of energetic particles responsible for the soft errors. Although numerous particle detection solutions have been reported, very few works address the on-chip particle detectors suited for the self-adaptive fault-tolerant microprocessor systems for space missions. This work reviews the state-of-the-art particle detectors, with emphasis on two solutions for the self-adaptive systems: particle detector based on embedded SRAM and particle detector based on pulse-stretching inverters.**

Keywords—Soft errors, particle detectors, self-adaptive fault tolerance

I. INTRODUCTION

The soft errors represent one of the most critical sources of failures in integrated circuits (ICs) employed in space missions. They are manifested as bit flips in storage elements (flip-flops, latches, and SRAM cells), also known as Single Event Upsets (SEUs). These events occur when an energetic particle hits a storage element and deposits sufficient charge to alter the stored logic value. Alternatively, the particle-induced voltage glitch in combinational logic, known as Single Event Transient (SET), can cause a soft error if it propagates through the logic path and is captured by a storage element.

As a result of Solar Particle Events (SPEs), the Soft Error Rate (SER) of an IC, i.e. the number of soft errors induced in a given time interval, can increase by several orders of magnitude [1]. Along with the SER variation due to radiation exposure in space, the downscaling of CMOS technologies has led to the exponential increase of the system SER [2]. This is primarily the result of dramatic increase in the number of on-chip elements with every new technology generation. Although the memory and sequential elements are dominant contributors to the overall SER because they occupy the largest area of a complex IC, the impact of combinational logic has increased significantly with the operating frequencies in the GHz range and the decrease of supply voltage and logic depth [3]. Therefore, the design of ICs for space applications requires special measures to mitigate the soft errors, i.e. to minimize the overall SER.

Besides the need for radiation hardness, the low energy consumption is also an essential design requirement for the space-borne electronics, because the energy resources in space are very limited. However, the reliability requirements are usually in conflict with the energy consumption constraints. For example, the reduction of the supply voltage decreases the energy consumption, but increases the system SER. Moreover, the fault tolerance is traditionally based on the hardware redundancy, which increases the energy consumption. Thus, the trade-off between fault tolerance and energy consumption is a major goal in the design process. A cost-effective approach to accomplish this is through the self-adaptive functionality - by adapting the operating modes of the system to the application and environmental conditions [4 – 7].

A typical example of a self-adaptive system is a multi- or many-core processor. The multiprocessing platforms have been introduced to overcome the processing limitations caused by the saturation of clock frequency with the technology scaling. In the past few years, the multi- and many-core systems have gained increased interest for space missions due to the increasing demand for the on-board real-time data processing [4]. By coupling the processing cores into various configurations, the trade-off between performance, energy consumption and fault tolerance can be maintained dynamically, thus extending the lifetime of the system. Depending of the radiation intensity in space and the application requirements, the fault tolerance mechanisms such as supply voltage and frequency scaling, Double Modular Redundancy (DMR) and Triple Modular Redundancy (TMR) can be implemented [4, 5].

It is necessary to monitor the radiation level during run-time to allow for dynamic configuration of the fault tolerance mechanisms. This is performed with the specially designed particle detectors, which operate on the principle of detecting the induced SETs or SEUs. Various types of semiconductor-based particle detectors for space applications exist and can be grouped into five main classes: (i) current detectors [8 – 15], (ii) acoustic wave detectors [16, 17], (iii) diode-based detectors [18 – 22], (iv) SRAM-based detectors [23 – 30], and (v) 3D NAND flash detectors [31 – 33]. Each type of particle detector has advantages as well as disadvantages, which are discussed in more detail in the following Sections.

The particle detectors for self-adaptive fault-tolerant systems must satisfy several requirements. First, the detectors should be sensitive to a wide range of particle energies and provide the

information on the particle flux, since the system SER increases linearly with the flux [34]. It is also important to monitor the variation of particle's Linear Energy Transfer (LET), because higher LET may result in multiple SEUs and longer SETs, and consequently in higher SER. The detectors should have fast response (low latency) and be robust to false alarms generated by other noise sources. Furthermore, the detectors should be integrated in the target chip to enable the in-situ monitoring of radiation exposure, and the readout electronics should introduce as low area and power overhead as possible.

However, none of the reported particle detectors [8 – 33] can satisfy all aforementioned requirements. Hence, there is a strong need for alternative solutions that can provide low-cost but accurate on-chip particle detection. Motivated by these goals, we have proposed two particle detection solutions: (i) a particle detector based on embedded SRAM [35] and (ii) a particle detector based on custom-sized pulse-stretching inverter chains [36, 37]. The preliminary evaluation has confirmed that both proposed solutions offer promising advantages over the state-of-the-art particle detectors in terms of the requirements for the self-adaptive fault-tolerant systems.

The rest of the paper is organized as follows. In Section II, the state-of-the-art particle detectors are briefly described. A concept of particle detection with embedded SRAM is presented in Section III, and in Section IV the particle detection with the pulse-stretching inverters is discussed. The comparison of the proposed solutions with the existing ones, in terms of the main performance metrics, is given in Section V. In Section VI, an example of a self-adaptive fault-tolerant multiprocessing system with particle detector is presented. The main directions for future work are outlined in Section VII.

II. State-of-the-Art Particle Detectors

A. Current Detectors

As energetic particles induce the current pulses in the target semiconductor device, the use of current sensors is a common approach for detecting these events. A simple design of a current sensor for detection of energetic particles was proposed in [8]. This sensor was connected to the power supply rail of SRAM. However, it was not suitable for detection of particle strikes in combinational logic because of the difficulty to differentiate the signal induced by a particle from the normal signal. An improved current sensor, known as the Bulk Built-in Current Sensor (BBICS), was proposed in [9, 10]. Instead of connecting to supply rail, BBICSs are connected to the bulk terminal of respective transistors. Separate BBICSs are needed for PMOS and NMOS transistors. When the bulk current exceeds the threshold level, the sensor generates a flag signal. The simplest structure of BBICS is composed of three transistors, as illustrated in Figure 1, but more sophisticated versions are more precise and reliable [11 – 13].

The major advantage of BBICSs is that they can provide the information on the faulty location, since they are connected directly to the target circuit. This enables to activate the error correction mechanisms only within the affected sub-circuit. It is not necessary to connect a sensor to each transistor, but one sensor can be utilized to monitor tens or thousands of transistors [12, 13]. This can be used as a guideline in planning the number and spatial distribution of BBICS on a chip, in order to achieve high detection efficiency with optimal number of sensors.

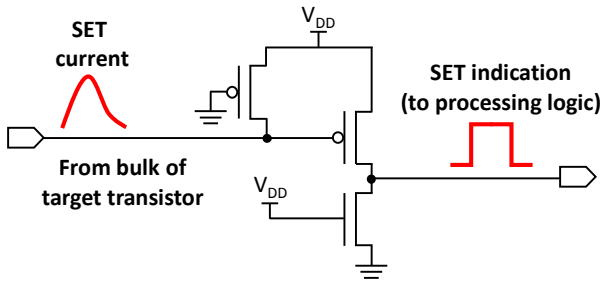


Figure 1: A simple BBICS design [10]

Nevertheless, the application of BBICS is associated with certain limitations. The key disadvantage of reported BBICS implementations is that only the particle strikes in the target circuit can be detected, but the information on the particle flux cannot be obtained directly. As the BBICSs are distributed across the chip, it is necessary to implement additional logic for collecting the data from all sensors and calculating the error rate from which the particle flux can be determined. However, there are no reports on any such implementations. In addition, the laser experiments performed on one version of the current detector [14] have revealed that BBICS sensitivity deteriorates with the increasing number of monitored transistors. A possible improvement by using the triple-well CMOS has been proposed [15], but this is not applicable to technologies with one or two wells. Moreover, as the BBICSs are connected to the target logic, they may be prone to other noise sources (e.g. substrate noise), which could lead to the triggering of false alarms.

B. Acoustic Wave Detectors

The monitoring of soft errors with acoustic wave detectors has been proposed in [16, 17]. Namely, a particle strike can generate the intense shock (acoustic) wave, which propagates through the substrate of the target circuit. For detecting such waves, a cantilever-like structure as depicted in Figure 2 can be used [17]. It acts as a capacitor and the particle strikes can be detected by measuring the capacitance change of the gap in cantilever. For this purpose, the mixed-signal processing logic is needed.

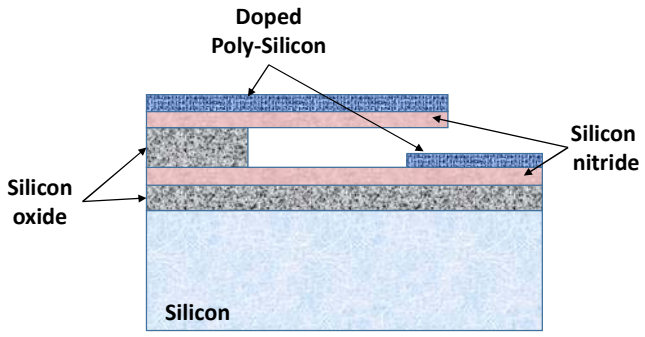


Figure 2: Cross-section of cantilever structure [17]

The solution proposed in [17] can be fabricated in CMOS technology, allowing easy integration into standard ICs. A single cantilever structure occupies the area of around 1 μm^2, which is roughly the area of an SRAM cell in 45 nm CMOS technology [17]. In order to achieve a sufficiently large sensing

area, a mesh implementation of multiple detectors is required. Proper dimensioning of the acoustic wave detector and choice of the appropriate number of detectors for the target chip is essential for achieving high sensitivity to particle strikes. Detailed guidelines for choosing the detector dimensions and for calibrating the detector are given in [17].

Like the BBICS, the acoustic wave detectors enable to detect the exact location of soft errors. This is achieved based on the relative time difference of arrival of acoustic wave for different detectors, using the algorithm given in [17]. However, the main drawback of the acoustic wave detectors is that their functionality has not been verified in practice yet. As these detectors need to be distributed across the chip, like the BBICS, they can provide the local detection of particle strikes, but for measuring the particle flux and LET is necessary to employ more complex processing circuitry.

C. *Diode Detectors*

The *p-n* junction (diode-based) detectors are one of the most widely used types of particle detectors. They are available in various implementations such as strip detectors, active pixel detectors and scintillator-coupled photodiodes [18 – 22]. In all implementations, the detectors are operated in reverse bias to achieve the minimum leakage current and maximum depletion layer width, thus ensuring the high detection efficiency. Radiation can induce either continuous or pulsed current in the detector, depending on the radiation intensity. Measuring the induced current enables to acquire the complete information on the radiation exposure and determine with high accuracy the induced charge, particle LET, flux, and energy spectra.

However, the use of diode-based detectors for triggering the dynamic fault tolerance mechanisms in a target IC may be too costly because different technologies have to be combined. The diodes are usually not manufactured in the same technology as the target system, which makes it challenging to integrate them on the same chip. Moreover, the need for mixed-signal processing increases the overall cost and complexity of the system. A typical structure of processing logic for a single diode is composed of a preamplifier, a pulse shaper, an analog-to-digital converter and a processor, as illustrated in Figure 3. As the practical implementations may be composed of hundreds or thousands of diodes on the same substrate, the hardware and power overhead due to the processing logic may be too high.

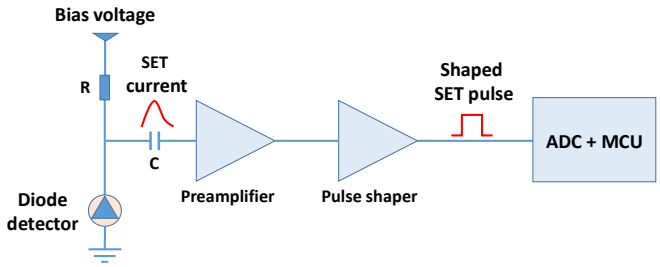


Figure 3: A processing channel for diode-based detector

D. *SRAM Detectors*

The use of commercial or custom-designed SRAMs as particle detectors, implemented as stand-alone ICs, has proven to be a very useful solution for soft error monitoring in various terrestrial and space applications [23 – 30]. The operation principle is based on counting the number of particle-induced SEUs in SRAM cells. When a particle hits a sensitive transistor within the cell, and deposits the energy exceeding the critical charge, the respective logic state will be changed from 0 to 1 or vice-versa. In general, the sensitivity is proportional to the size of SRAM (number of SRAM cells). The most common implementations employ the six-transistor (6T) SRAM cells as illustrated in Figure 4. Based on the detected number of SEUs and the cross-section of SRAM obtained experimentally, the particle flux can be calculated. The SEUs are detected and corrected using some of the well-known Error Detection and Correction (EDAC) mechanisms and memory scrubbing. The response time of the SRAM-based detector is determined by the scrubbing rate, which is on the other hand defined by the clock frequency of the system.

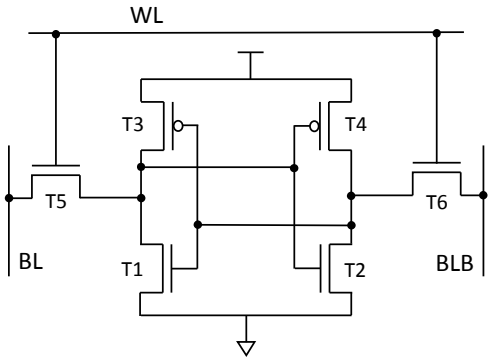


Figure 4: 6T SRAM cell

The main advantage of SRAM detectors is simple operating principle, no need for analog processing and possibility of manufacturing in the same technology as the standard ICs. However, this approach has several limitations. In the stand-alone implementations [23 – 30], the area overhead due to the EDAC logic may be too large. From functional point of view, the EDAC techniques suffer from the limitation in the number of detectable and correctable errors, which may lead to the error accumulation because of multiple upsets. Moreover, due to scrubbing the response of SRAM detectors may be slower compared to other solutions.

While most reported SRAM-based solutions have been used only for flux measurement, it is also possible to measure LET with custom-designed SRAM detectors. A solution presented in [29, 30] uses the custom-designed SRAM detector which generates SET pulses in response to particle strikes, and the analog processing logic is used to amplify the pulses and then measure their amplitude. Based on measured pulse amplitude, the energy and LET of incident particles can be determined. In this case, all cells are connected in parallel, so that a particle strike induces a pulse, which can propagate to the output. The solution is designed as a standalone spectrometer and as such is not suitable for the on-chip integration.

E. *3D NAND Flash Detectors*

Recently, the use of 3D NAND flash memory with floating gate transistors as a heavy-ion detector has been proposed [31]. Although the detectors based on floating gate transistors have been used for total dose measurement [32, 33], the work [31] is the first to verify the applicability of this concept for detection

of energetic particles. The operating principle relies on measurement of the threshold voltage shift of floating gate transistors due to the charge deposited by the incident particles. This allows to not only measure the error rate, but also the particle LET. In addition, due to the 3D structure of memory, the angle of incidence can be estimated. Furthermore, the 3D structure allows to differentiate between the errors induced by incident particles and the errors due to electrical noise, as well as to differentiate between single and multiple upsets.

Although the 3D NAND flash with floating gate transistors is a promising solution with substantial benefits over other detectors such as SRAM, the main limitation currently is difficulty in integrating it in the target chip. Due to the 3D structure and floating gate technology, this approach may be too complex for integration into a conventional planar CMOS IC designed with standard design tools. In addition, the processing electronics may be complex and costly because it is necessary to measure precisely the change of the threshold voltage of floating gate transistors, which requires the use of analog processing circuitry and analog-to-digital converters.

III. Embedded SRAM as a Particle Detector

As alternative to the conventional stand-alone SRAM-based particle detectors described in previous Section, we have proposed the use of embedded SRAM as a particle detector [35]. The idea is to employ the standard on-chip SRAM memory as a particle detector in parallel to its normal data storage function. The detection principle is the same as for the stand-alone SRAM detectors discussed in previous Section, i.e. the SEUs detected in SRAM cells are counted and, from this information, the particle flux can be determined. A similar approach, based on Block Random Access Memory (BRAM) in FPGA was introduced in [1]. However, in contrast to all previous solutions, the proposed embedded SRAM monitor incurs significantly less area overhead because the available on-chip resources are utilized for particle detection. An important feature of the proposed solution is the capability to detect the permanent faults in SRAM cells. This is essential for maintaining the accurate SEU measurements in long-term missions, where the permanent errors occur due to the gradual device wear-out.

Figure 5 shows the block diagram of the embedded SRAM with the support for particle detection. It consists of a Synchronous SRAM (SSRAM) with five 512k × 8-bit asynchronous SRAM blocks, a Control Unit, a Scrubbing module and an EDAC module. Four memory blocks are used for data storage and particle detection, while one block is allocated for storing the 7-bit EDAC syndrome computed on each 32-bit word written in the other four memory blocks. Thus, the user sees effectively a 16-Mbit memory device. The memory blocks are based on the 6T SRAM cell shown in Figure 4. Each read, write or scrubbing cycle uses the EDAC module and involves the access to 32 bits selected by a 19-bit address. As the sequential logic in the Control Unit, EDAC and Scrubbing modules is inherently sensitive to SEUs, the Triple Modular Redundancy is applied to all flip-flops.

The functions of EDAC and Scrubbing modules is to protect the memory cells against SEUs and detect the single and double bit errors as well as permanent faults in each memory word. The built-in EDAC module performs the Single-Error Correction and Double-Error Detection (SEC-DED) with (39, 32) HSIAO code. The HSIAO code was chosen because it provides fast coding/decoding with low hardware overhead. The three 8-bit Error Counters are integrated in the Control Unit to count the single and double bit errors and permanent faults. Any error that cannot be corrected by EDAC is considered as a permanent error. A detailed description of the algorithm for detection of single, double and permanent faults can be found in [35]. The number of detected faults is stored in the Register File to avoid duplicate counting of the double and permanent faults.

The primary role of the Scrubbing module is to avoid accumulation of radiation-induced soft errors. The scrubbing module periodically reads the memory locations when the chip is in the idle state. When the error is detected, the EDAC procedure is activated. The scrubbing is entirely autonomous and transparent for the user, which means that the user can access the SSRAM even if the scrubbing procedure is in progress. The scrubbing rate (response time) can be configured by the user, but it is limited by the operating frequency.

The proposed SRAM monitor has been designed in IHP's 130 nm bulk CMOS technology with the nominal supply voltage of 1.2 V. The recommended operating frequency for this design is 50 MHz. For this frequency, the minimum scrubbing rate is 42 ms. With respect to the total area of the design, the introduced area overhead is less than 1%, while the power overhead is even less, below 0.1 %. The elements contributing to the area and power overhead are the Error Counters and the Register File, while all other hardware resources are employed in any rad-hard SRAM. A detailed discussion of the synthesis results can be found in [33].

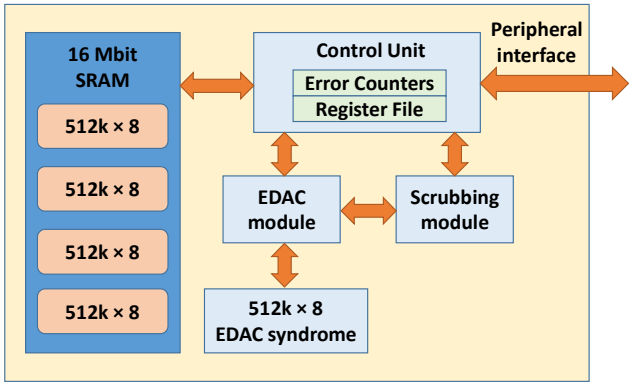


Figure 5: Embedded SRAM as a particle detector

However, the main issue with the on-chip SRAM used as a particle detector is the limited sensitive area. In general, the SRAM should be as large as possible to obtain sufficiently large sensitive area and thus ensure the high probability of particle detection. Previous studies have shown that the SRAM with the capacity of several Gbit is needed for sufficient sensitivity, but these solutions are based on standalone SRAMs. When the data-storage SRAM within the target chip is used as a particle detector, its sensitive area will be constrained by the application requirements. The size of the on-chip SRAM is usually limited to tens or hundreds of Mbits. To acquire statistically relevant number of SEUs with a smaller on-chip SRAM, it is necessary to employ longer detection intervals.

IV. Pulse-Stretching Inverter Chain as a Particle Detector

The application of custom-sized pulse-stretching inverter chains as particle detectors has been proposed in our previous work [36 – 38]. The idea is to measure the SET count rate and SET pulse width variations. Thereby, the particle flux can be determined in terms of the SET count rate, while the LET variations can be determined in terms of the SET pulse width variations. It is important to note that this solution cannot measure the exact SET pulse width, but only to sort the detected SET widths into several distinct ranges. This is because the digital processing, as a simple and low-cost alternative to the analog processing in diode detectors, has been chosen in this case. Nevertheless, the information on the SET count rate and SET pulse width variation is sufficient for the target self-adaptive fault-tolerant systems.

A basic sensing element consists of two inverters connected in series, and this configuration is denoted as a Pulse-Stretching Cell (PSC). By setting the fixed logic level at the input of PSC, two transistors will always be in on state while the other two will be in off state. The off-state transistors are sensitive to particle strikes while the on-state transistors act as restoring elements (provide the current to compensate the particle-induced charge). The PSCs have skewed sizing, i.e. in one inverter, the PMOS transistor has larger channel width than NMOS transistor, while in the other inverter the NMOS transistor has larger channel width than the PMOS transistor. To achieve sufficiently large sensing area of a PSC and thus increase the probability of particle strikes, the off-state transistors should have as large channel width as possible. On the other hand, to decrease the restoring current and thus increase the sensitivity, the on-state transistors should have small channel width and large channel length. Furthermore, the skewed sizing ensures that the SET pulse induced in the PSC is stretched as it propagates through the chain. Therefore, even the low energy particles can results in observable SETs. A detailed description of the transistor sizing for the PSC can be found in [36 – 38].

Using a single PSC is generally not sufficient because the two sensitive transistors still have quite small sensing area. The sensing area can be increasing by connecting an appropriate number of PSCs in series or in parallel. In serial configuration, two detector versions are possible: (i) a long chain of PSC or (ii) a number of shorter PSC chains connected by an OR tree. In parallel configuration, the number of PSC that can be connected in parallel is limited due to the loading effects, i.e. a large number of PSCs in parallel reduces the sensitivity. Thus, a number of arrays made of PSCs connected in parallel have to be employed, as illustrated in Figure 6. The serial configuration is suitable only for measuring the SET count rate because the SET pulse width changes very little over a wide range of LET. On the other hand, the parallel configuration enables to capture both SET count rate and SET pulse width variation.

Both configurations have been evaluated with SPICE simulations, using the bias-dependent current model to simulate SET effects. The analysis was performed for IHP's 130 nm CMOS technology. It was shown that with large off-state transistors and small on-state transistors, the threshold LET is below 0.2 MeVcm^{2}mg^{-1}. This is lower than the threshold LET of standard logic cells and the LET of common particles encountered in space. For the parallel configuration depicted in Figure 6, the SET pulse width changes by approximately 550 ps in the LET range from 1 to 100 MeVcm^{2}mg^{-1}, whereby the maximum SET width is in the order of several ns. On the other hand, in the serial configuration the output SET pulse width is directly proportional to the number of cascaded PSCs and can be from hundreds of ps to hundreds of ns. Therefore, the pulse-stretching detector is expected to have faster response than the SRAM-based detectors. Moreover, the pulse-stretching detector is immune to error accumulation because of the transient nature of SET effects.

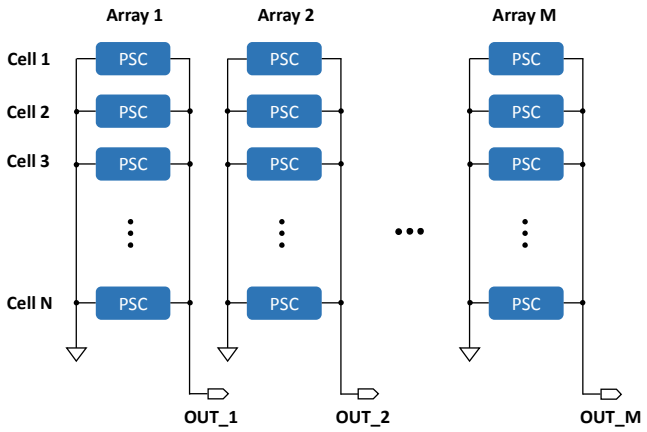


Figure 6: Particle detector based on arrays of PSCs connected in parallel

Figure 7 illustrates the general architecture of the processing logic for the particle detector composed of parallel arrays of PSCs illustrated in Figure 6. The outputs of all pulse-stretching arrays are connected to a standard OR-tree to obtain a single output, which is then interfaced to the processing logic. An SET induced in any array will propagate to the output of the OR-tree and then further through the SET filters and respective SET counters. The SET filters allow the propagation of SET pulses within predefined pulse width ranges. Thus, the corresponding counters store the number of detected SETs with the predefined pulse widths. The control unit reads the current state of all counters, stores the acquired results in register file, and resets periodically the counters. The standard hardening measures can be applied to the processing logic.

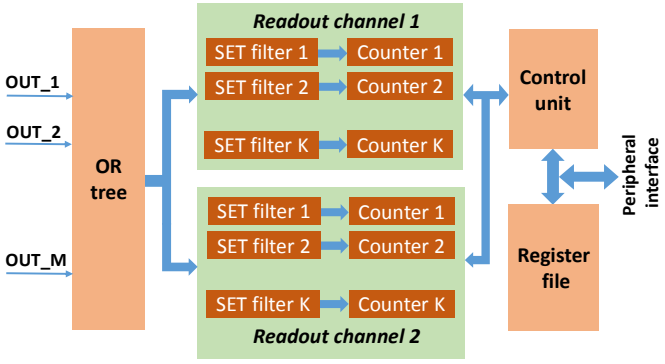


Figure 7: Readout circuit for pulse-stretching particle detector based on parallel PSC arrays

V. Comparison of Particle Detectors

Based on the published results, a comparative analysis of the discussed particle detectors in terms of six performance metrics is presented in Table 1. The advantages and dis-advantages of each detector type should be carefully considered in selecting the appropriate detector for a particular application. For the space applications where the self-adaptive dynamic fault tolerance is required, it is important to integrate the particle detector in a single chip with the target system. This enables to sense directly the radiation to which the target system is exposed. In that context, the two detectors proposed in our previous work offer essential advantages over the state-of-the-art solutions.

The main advantage of embedded SRAM-based detector over all other solutions is that it serves as a standard data storage memory in a target system. This results in negligible area and power overheads since the existing on-chip resources are used for particle detection. As a result, the cost of implementation is lower compared to other solutions. In addition, the possibility of detecting the permanent errors is important advantage over all other detectors, as none of the existing solutions supports this functionality.

On the other hand, the particle detector based on the pulse-stretching inverter chains offers the possibility to measure the LET variation, which is possible also with diode and 3D NAND flash detectors. However, compared to these detectors, the pulse-stretching detector employs simple digital processing logic, which minimized both the area and power overheads and thus the overall cost. Moreover, the immunity to multiple errors is an advantage over the conventional SRAM detectors.

Table 1: Comparison of particle detectors

Type of detector	Probability of false alarms	Complexity of readout logic	Hardware/ power overhead	Sensitivity to multiple errors	Ability to monitor LET variation	Additional functions
Built-in current detector	Moderate	Low	Medium	No	No	No
Acoustic wave detector	Moderate	Moderate	Medium	No	No	No
Diode detector	Low	High	High	No	Yes	No
Stand-alone SRAM detector	Low	Low	High	Yes	No	No
3D NAND flash detector	Low	High	High	No	Yes	No
Embedded SRAM detector	Low	Low	Low	Yes	No	Data storage and detection of permanent errors
Pulse-stretching detector	Low	Low	Medium	No	Yes	No

VI. Application Scenario: Self-Adaptive Quad-Core Processing System

To illustrate the operation of a self-adaptive fault-tolerant multiprocessing system with a built-in particle monitor, we have chosen a quad-core platform as an example.

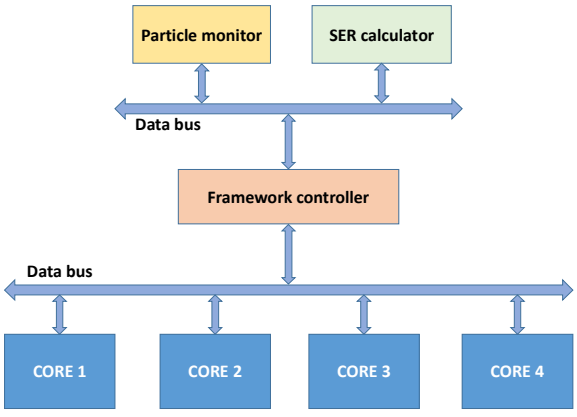


Figure 8: General architecture of a self-adaptive quad-core processing platform

The block diagram of the self-adaptive fault-tolerant quad-core system is illustrated in Figure 8. The basis of the system is the *Waterbear* framework controller, which enables to select the three main operating modes [4]:

- **High performance mode**: the multiprocessor operates as a common multiprocessor, i.e. each core executes its own task. This mode is selected according to the application requirements.
- **Destress (or low power) mode**: a single core is operating while the others are clocked- or powered-off to reduce aging and save energy. The anti-aging technique known as Youngest-First Round-Robin (YFRR) core gating [39] is employed to de-stress the operating core by transferring the workload to a resting core. This is done periodically and with the help of special aging monitors embedded in each core [40].
- **Fault tolerance mode**: the processing cores are coupled into various fault-tolerant configurations such that they are executing each instruction simultaneously. The voting is performed in each cycle by a voter unit which initiates the actions (e.g. interrupt request or reset) when the mismatch between the core outputs is detected [4].

During operation under radiation exposure in space, the particle monitor generates the information on the SET or SEU count rate and LET (if the chosen detectors supports the LET measurement). The SER calculator processes the information from the particle detector to determine the real-time SER variation of the multi-core system. The SER calculator can be extended with a hardware accelerator module for prediction of SPEs, at least one hour in advance, based on the supervised machine learning, as detailed in [41, 42]. This functionality allows for the early detection of the increasing radiation level (which results in increased SER) and timely activation of the respective fault-tolerant mechanisms.

Based on the measured or predicted SER, various fault-tolerant solutions can be applied at the core level to achieve the radiation hardness, such as:

- **Supply voltage and frequency scaling:** By increasing the supply voltage and decreasing the operating frequency, the SER is reduced at the cost of increased energy consumption and reduced processing speed. This approach can provide limited improvement in SER, which could be valuable at low and medium level radiation levels. In this case, either all cores are engaged in parallel processing or some of them may be switched off.
- **Double modular redundancy (DMR):** In this mode, the four cores can be divided into two pairs of DMR cores, such that the system is essentially operating as a dual-core system with enhanced fault tolerance. This approach is useful under medium radiation exposure, but the drawback is the reduced processing speed.
- **Triple modular redundancy (TMR):** In this mode, three cores are coupled into a TMR configuration while the fourth core is powered off. Thus, the system operates as a single core with the highest level of protection under high radiation levels. The main drawback of this approach is the reduced processing speed because all cores perform the same task.

The concept illustrated in Figure 8 is flexible and can be adopted to a larger number of cores with minor modifications of original design. To accommodate the particle detector and SER calculator, the original framework controller design requires the addition of an interface for processing the data from the added modules. If the platform is applied to a many-core system, the DMR and TMR configurations can be implemented on multiple groups of processing cores. For example, in an assumed 8-core system would be possible to have two TMR blocks operating as a dual-core processor, thus achieving the high level of fault tolerance and at the same time providing enhanced performance. This concept has been verified on an 8-core 32-bit chip demonstrator designed and manufactured in IHP's 130 nm bulk CMOS technology [5].

The main benefit of the multi-core approach in terms of fault tolerance is that the inherent hardware redundancy is used as a basis for achieving the fault tolerance. The processing cores are considered as redundant only in the fault tolerance mode while in the high performance mode they are employed for multiprocessing. As a result, the area overhead is minimal and is related only to the additional logic that is needed for selecting the fault tolerance modes.

VII. Conclusion and Future work

In this work, the comparative analysis of several solutions for detection of energetic particles responsible for soft errors in integrated circuits is presented. The comparison was performed based on the requirements for the online particle detection in the self-adaptive fault-tolerant systems for space applications. Beside the five state-of-the-art semiconductor particle detectors (diode-based, SRAM-based, bulk built-in current, acoustic wave and 3D NAND flash detectors), we have introduced the two detector concepts which have resulted from our ongoing research – the embedded SRAM-based detector and the pulse-stretching detector. The comparative analysis has shown that the two proposed solutions have remarkable advantages over the existing particle detectors regarding the self-adaptive fault tolerance applications.

Future work will be directed towards experimental validation of the two proposed particle detectors. To this end, it is necessary to conduct the irradiation campaign with the detector prototypes, in order to calibrate their response and determine the optimal design specifications.

Acknowledgment

This work was done in the framework of project REDOX (funded by the German Research Foundation DFG under the grant agreement No. KR 3576/29-1), and project RESCUE (funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 722325).

References

[1] R. Glein et al., "Detection of Solar Particle Events inside FPGAs," in Proc. European Conference on Radiation Effects on Components and Systems (RADECS), 2016.

[2] E. Ibe, H. Taniguchi, Y. Yahagi, K.i. Shimbo, and T. Toba, "Impact of scaling on neutron-induced soft error in SRAMs from a 250 nm to a 22 nm design rule," IEEE Transactions on Electron Devices, 2010.

[3] N. N. Mahatme et al., "Impact of Technology Scaling on the Combinational Logic Soft Error Rate," in Proc. IEEE International Reliability Physics Symposium (IRPS), 2014.

[4] A. Simevski, O. Schrape, C. Benito, M. Andjelkovic, M. Krstic, "PISA: Power-Robust Microprocessor Design for Space Applications," in Proc. International Symposium on Online Testing and Robust System Design (IOLTS), 2020.

[5] M. Krstic, A. Simevski, M. Ulbricht, S. Weidling, "Power/Area-Optimized Fault Tolerance for Safety Critical Applications," in Proc. International Symposium on Online Testing and Robust System Design (IOLTS), 2018.

[6] C. Bolchini et al., "Self-Adaptive Fault Tolerance in Multi-/Many-Core Systems," Journal of Electronic Testing, 2013.

[7] A. Jacobs et al., "Reconfigurable Fault Tolerance: A Comprehensive Framework for Reliable and Adaptive FPGA-Based Space Computing," ACM Transactions on Reconfigurable Technologies and Systems, 2012.

[8] B. Gill et al., "An Efficient BICS Design for SEUs Detection and Correction in Semiconductor Memories," in Proc. Design, Automation and Test in Europe Conference (DATE), 2005.

[9] E. H. Neto, I. Ribeiro, M. Vieira, G. Wirth, and F. L. Kastensmidt, "Evaluating Fault Coverage of Bulk Built-in Current Sensor for Soft Errors in Combinational and Sequential Logic," in Proc. Symposium on Integrated Circuits and Systems Design (SBCCI), 2005.

[10] G. Wirth, "Bulk Built-In Current Sensors for Single Event Transient Detection in Deep-Submicron Technologies," Microelectronics Reliability, 2008.

[11] H.-B. Wang, R. Liu, L. Chen, J.-S. Bi, M.-L. Li, Y.-Q. Li, "A Novel Built-in Current Sensors for N-Well SET Detection," Journal of Electronic Testing, 2015.

[12] R. Possamai Basstos, L. Acunha Guimaraes, F. Sill Torres, L. Fesquet, "Architectures of Bulk Built-in Current Sensors for Detection of Transient Faults in Integrated Circuits," Microelectronics Journal, 2018.

[13] R. Possamai Basstos et al., "Assessment of On-Chip Current Sensor for Detection of Thermal-Neutron Induced Transients," IEEE Trans. on Nuclear Science, 2020.

[14] Z. Zhang et al., "A Bulk Built-In Voltage Sensor to Detect Physical Location of Single-Event Transients," Journal of Electronic Testing, 2013.

[15] J. M. Dutertre et al., "Improving the Ability of Bulk Built-In Current Sensors to Detect Single Event Effects by Using Triple-Well CMOS," Microelectronics Reliability, 2014.

[16] E. Hannah, "Cosmic Ray Detectors for Integrated Circuit Chips," US Patent US7309866B2, 2007.

[17] G. Upasani, H. Vera, A. Gonzales, "A Case for Acoustic Wave Detectors for Soft-Errors," IEEE Trans. on Computers, 2016.

[18] W. S. Wong et al, "Introducing Timepix2, A Frame-Based Pixel Detector Readout ASIC Measuring Energy Deposition and Arrival Time," Radiation Measurements, 2020.

[19] M. Havranek et al., "MAPS Sensor for Radiation Imaging Designed in 180 nm SOI CMOS Technology," Journal of Instrumentation, 2018.

[20] C. I. Underwood et al., "Radiation Environment Measurements with the Cosmic Ray Experiments On-Board the KITSAT-1 and PoSAT-1 Micro-Satellites," IEEE Trans. on Nuclear Science, 1994.

[21] S. Roy et al., "Plastic Scintillator Detector Array for Detection of Cosmic Ray Air Shower," Nuclear Insturments and Methods in Physics Research A, 2019.

[22] S. Kasahara et al., "Application on Single-Sided Silicon Strip Detectors for Energy and Charge State Measurements of Medium Energy Ions in Space," Nuclear Instruments and Methods in Physics Research A, 2009.

[23] R. Harboe-Sorensen et al., "Design, Testing and Calibration of a Reference SEU Monitor System," in European Conference on Radiation Effects on Components and Systems (RADECS), 2005.

[24] G. Tsiligiannis et al., "An SRAM Based Monitor for Mixed-Field Radiation Environments," IEEE Trans. Nuclear Science, 2014.

[25] L. Dilillo, A. Bosser, V. Gupta, F.Wrobel, F. Saigne, "Real-Time SRAM Based Particle Detector," in Proc. Intern. Workshop on Advances in Sensors and Interfaces (IWASI), 2015.

[26] K. S. Ytre-Hauge et al., "Design and Characterization of SRAM-Based Neutron Detector for Particle Therapy," Nuclear Instruments and Methods in Physics Reasearch A, 2015.

[27] R. Secondo et al., "Embedded Detection and Correction of SEU Bursts in SRAM Memories Used as Radiation Detectors," IEEE Trans. on Nuclear Science, 2016.

[28] J, Prinzie et al., "An SRAM-Based Radiation Monitor with Dynamic Voltage Control in 0.18 μm CMOS Technology," IEEE Trans. on Nuclear Science, 2019.

[29] E. G. Stassinopoulos, C. A. Stauffer, G. J. Brucker, "Miniature high-LET radiation spectrometer for space and avionics applications," Nuclear Instruments and Methods in Physics Research A, 1998.

[30] E. G. Stassinopoulos, J. L. Barth, C. A. Stauffer, "Measurement of Cosmic Ray and Trapped Proton LET Spectra on the STS-95 HOST Mission," IEEE Trans. on Nuclear Science, 2017.

[31] M. Bagatin et al., "A Heavy-Ion Detector Based on 3-D NAND Flash Memories," IEEE Trans. on Nuclear Science, 2020.

[32] E. Pikhay, Y. Roizin, U. Gatti, C. Calligaro, "Re-usable 180 nm CMOS Dosimeter Based on Floating Gate Transistor," in Proc. IEEE Intern. Conference on Electronic Circuits and Systems (ICECS), 2016.

[33] M. Brucoli et al., "A Complete Qualification of Floating Gate Dosimeter for CERN Applications," in Proc. European Conference on Radiation Effects on Components and Systems (RADECS), 2016.

[34] P. Hazucha, C. Svensson, "Impact of CMOS Technology Scaling on the Atmospheric Neutron Soft Error Rate," IEEE Trans. on Nuclear Science, 2000.

[35] J. Chen, M. Andjelkovic, A. Simevski, Y. Li, M.Krstic, "Design of SRAM-Based Low-Cost SEU Monitor for Self-Adaptive Multiprocessing System," in Proc. Euromicro Conference on Digital System Design (DSD), 2019.

[36] M. Andjelkovic. M. Veleski, J. Chen, A. Simevski, M. Krstic, R. Kraemer, "A Particle Detector Based on Pulse Stretching Inverter Chain," in Proc. IEEE Intern. Conference on Electronic Circuits and Systems (ICECS), 2019.

[37] M. Andjelkovic, J. Chen, A. Simevski, M. Krstic, R. Kraemer, "Monitoring of Particle Flux and LET with the Pulse Stretching Inverters," in Proc. European Conference on Radiation Effects on Components and Systems (RADECS), 2020. (Accepted paper)

[38] M. Andjelkovic, M. Krstic, R. Kraemer, "Study of the Operation and SET Robustness of a CMOS Pulse Stretching Circuit," Microelectronics Reliability, 2018.

[39] A. Simevski, R. Kraemer, M. Krstic, "Increasing Multiprocessor Lifetime by Youngest-First Round-Robin Core Gating Patterns," in Proc. NASA/ESA Conference on Adaptive Hardware and Systems (AHS), 2014.

[40] A. Simevski, R. Kraemer, M. Krstic, "Low-Complexity Integrated Circuit Aging Monitor," in Proc. International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS), 2011.

[41] J. Chen, T. Lange, M. Andjelkovic, A. Simevski, M. Krstic, "Prediction of Solar Particle Events with SRAM-Based Soft Error Rate Monitor and Supervised Machine Learning," Microelectronics Reliability, 2020. (Accepted paper)

[42] J. Chen, T. Lange, M. Andjelkovic, A. Simevski, M. Krstic, "Hardware Accelerator Design with Supervised Machine Learning for Solar Particle Event Prediction," in Proc. IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), 2020. (Accepted paper)

SEkey: A Distributed Hardware-based Key Management System

Matteo FORNERO
CINI Cybersecurity National Lab.
Turin, Italy
matteo.fornero@consorzio-cini.it

Nicolò MAUNERO
Politecnico di Torino
CINI Cybersecurity National Lab.
Turin, Italy
nicolo.maunero@polito.it

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

Antonio VARRIALE
B5 Labs Ltd.
Ta' Xbiex, Malta
av@blu5labs.eu

***Abstract*—Cryptography plays a key role in all the aspects of today cybersecurity and any cryptographic approach relies on cryptographic keys, i.e., series of bits that determine how a plain text is encrypted and decrypted, according to an agreed algorithm. The secrecy and security of an encryption key are thus crucial and fundamental: if the cryptographic key is compromised and known, everyone can decrypt a text encrypted according to the strongest encryption algorithm. As a consequence, several Key Management Systems (KMS) have been developed to easily support the management of cryptographic keys, whose number is constantly increasing, due to the amount of devices and communications that take place today, even in very restricted contexts. SEkey is a key management system developed targeting a distributed environment, where it is possible to identify a single central manager that acts as a Key Distribution Center (KDC) and many users that locally store and manage their own keys. Users, to a certain extent, can also work 'offline' without being always in direct communication with the central manager. SEkey is built leveraging the functionalities and physical properties of the SEcube™ Hardware Security Module (HSM). All the key values and critical information are stored inside the SEcube™ and never leave the device in clear, and all the cryptographic operations are performed by the SEcube™ itself. The guidelines provided by NIST where followed during the whole development process, guaranteeing all the most important security features and principles.**

I. Introduction

The increase in the number of connected devices, that has been taking place for several years now, is posing several security challenges by considerably enlarging the cyber attack surface. The quantity and quality of data exchanged every second among people and various devices is increasing at an exponential rate, making it mandatory to secure them.

Cybersecurity is a term that includes several concepts, but the fil rouge that connects them all is *cryptography*. As defined by NIST [4], cryptography is the discipline that embodies the principles, means and methods for the transformation of data in order to hide their semantic content, prevent their unauthorized use ore avoid their undetected modification. This data transformation process takes place through mathematical operations, more or less complex, that combine together the input data, usually referred to as cleartext, and the cryptographic key to obtain the modified data as output, what is usually referred to as ciphertext.

The cryptographic key is a parameter used in conjunction with a cryptographic algorithm that determines its spectrum of operation [7]. Drawing a parallel with everyday life, the role of a cryptographic key is similar to the key of a lock. Locking is like data encryption while unlocking is like data decryption and, just as in the case of a lock, also in cryptography protecting the key is of paramount importance: even in presence of the best encryption algorithm, if the cryptographic key is compromised and everyone knows it, then everyone can access the encrypted data.

Nowadays the amount of keys and the requirements for their security make it practically impossible to manage them by hand; for this reason the so-called *Key Management Systems* (KMS) were born, applications that aim to automate and simplify the management of cryptographic keys in highly complex contexts.

In this paper we present SEkey, a mixed hardware-software Key Management System, leveraging on the SEcube™[1] Hardware Security Module (HSM). SEkey is designed and developed having in mind a distributed ecosystem where each entity gets its own SEcube™ device. Each SEcube™ is in charge of securely storing all the encryption keys and of providing all the security primitives for securely managing keys and performing cryptographic operations. This allows to never expose the actual key value outside of the device when performing crypto operations. In addition, during the key distribution process, keys are over-ciphered with a unique key shared only by the administrator and the user that receive the update. Inside SEkey, two roles are available: the *security administrator* and the *user*. The former is in charge of distributing keys and synchronising all the SEcubes™, while the user passively uses its device for security purposes, everything related to the key management being automatically handled by the SEcube™ device.

The paper is structured as follows: the next section introduces a brief overview on the SEcube™ project and device. In the second section a brief analysis of the SOA on different types of KMS is proposed and the most important guidelines from the NIST for KMS development are reported. Then the implementation details and the most relevant features of SEkey

[1] https://www.secube.eu/

are presented, concluding with possible improvements and future works.

II. The SEcube™ Open Security Platform

The SEcube™ Open Security Platform [18] leverages on the functionalities of the SEcube™ SoC to provide a security-oriented open software and hardware platform. The SEcube™ SoC, developed by the Blu5 Group Company, includes three main cores:

- A STM32F4 microcontroller unit, equipped with an ARM Cortex-M4 processor.
- A reconfigurable hardware device (FPGA).
- An EAL 5+ certified Smart Card.

A 3D packaging of the three components and a set of custom technological solutions improve the resiliency to side-channel attacks [5] and to attempts to exfiltrate data from the device.

The SEcube™ platform is equipped with set of high-level APIs that abstract complex concepts of cybersecurity and cryptography [19], designed to ease the development of high security applications. Among the others, the open source libraries [8] include *SEfile* [9] and *SElink* [8], aimed at protecting data at rest and data in motion, respectively [20]. In particular, SElink provides a set of API that can be used to securely handle communications channels via end-to-end encryption, whereas SEfile provides a set of API for handling files in a secure way, allowing secure implementations of the most common system calls of the Posix Portable Operating System Interface and WIN32. These APIs are a simplified version of these system calls, not exposing all the functionalities provided by them, but managing internally all the security operations required to handle encrypted files.

III. Background

A. Key Management Systems Overview

Key Management Systems can be clustered according to different categories, including the way they are provided to the customers, the organization of the Key Distribution Center, and their key storage facilities.

According to the way a KMS is provided to the customers, four categories are mostly used: *software*, *virtual*, *appliance*, and *service* [6].

A *software KMS* is purely software-based and either implements its own protocol or is compliant with standard ones. The software runs on an Operating System (OS) that is hosting the KMS (typically, a server built by the customers to accommodate the KMS software).

A *virtual KMS* is a pre-installed virtual machine that runs the KMS software in a virtualised environment. The hardware where the VM runs is not shipped with the MKS and is under control and responsibility of the customers.

An *appliance KMS* is an integrated hardware-software solution. In this case both hardware and software are provided to the customer and they can be, for example, a server with certified hardware and software or a KMS running or leveraging on a hardware security module.

Type	Pros	Cons
Software	Wide compatibility	HW and OS provided by customer
	Runs on pre-existing hardware	Hardware may not be certified
	Runs on common OS	OS may not be certified
	Easy to fix and update	Usually weaker
Virtual	Wide compatibility	HW provided by customer
	Runs on pre-existing hardware	Hardware may not be certified
	Easy to run multiple installations	Usually weaker
	OS provided with the KMS	Virtualization overhead
	Easy to fix and update	
Appliance	HW and SW provided with the KMS	Lower flexibility
	Turnkey installation	Difficult to fix or update
	All-in-one solution	HW limitations
	Usually more secure	Usually more expensive
Service	No installation required	Keys stored in the cloud
	Easy to use	No physical control
	No local resources required	
	Flexible in terms of usage and payments	

TABLE I
Pros and cons of different kinds of KMS [6]

A *service KMS* is a cloud-based solution that can be used by the customers without the need of a specific hardware or infrastructure. This approach is also known as *KMS-as-a-service* and it is one of the most used solutions due to its flexibility and its migration capabilities.

Table I summarises pros and cons of each solution.

When categorised according to their *Key Distribution Center* (KDC), i.e., the entity responsible for distributing keys, key management systems are usually clustered as *distributed*, *centralized* and *decentralized* [14]. A *centralized KMS* is built around a single central entity that is in charge of managing the keys and distributing them to all the users. In a *distributed KMS* there is no single master entity and each user of the KMS manages her/his own keys and uses contributory key agreement protocols [2] to cooperate and contribute, with all other members of the group, to the creation of a shared key. In a *decentralized KMS* users are split into several smaller subgroups, each managed by an appointed manager who can, in turn, refer or not to a manager of the entire KMS.

With respect to the adopted key storage solution, KMS's are usually defined as *centralised* or *distributed* [10]. In the former case all the keys are stored by the master entity of the KMS that is in charge of providing secure storage for all of them, whereas in the latter one each user is in charge of storing her/his own keys in a secure way and should be provided with all the tools necessary to fulfil this requirement. An example of distributed KMS can be the Apple Secure Enclave Processor (SEP) [13] an isolated component from the main processor that provides secure storage for critical information, finger print, cryptographic keys, etc. but also cryptographic primitives for the main system.

B. NIST Recommendation

US NIST plays a key role in providing guidelines and recommendation for Key Management and KMS development [3], today widely and extensively adopted by the implementers

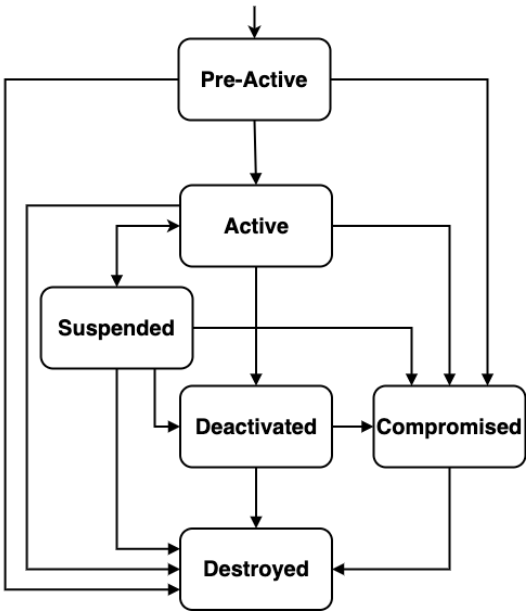


Fig. 1. Key State and Transactions

of KMS's worldwide. In the sequel we briefly recall some of the most significant issues pointed out in the NIST's documents.

Key life cycle: at any given point in time, a key is characterised by a specific *state* [3] that determines how the key can be used. Figure 1 shows the different states and the permitted transitions among states.

- *Pre-activation*: when a key is created it enters this state and it cannot be used until "activated".
- *Active*: in order to be used, a key must be in this state.
- *Suspended*: when in this state, the key cannot be used, but it can be activated again.
- *Deactivated*: the key can be used only to decrypt, but no longer to encrypt. When a key is replaced by a newer one, it is still needed for decrypting data encrypted by it.
- *Compromised*: this is a warning state. It means that the key is, or may be, compromised due to, for example, a data breach; the key can still be used for both encryption and decryption but with particular care. A compromised key cannot be reactivated.
- *Destroyed*: when in this state, the key is completely removed from the system.

Cryptoperiod is the time span during which a specific key can be used. This quantity is extremely important and it is strictly related to the security of a cryptographic key, the more a key is used, the more frequently it must be updated in order to lower the chance for it to be compromised.

Physical and logical access protection this is of paramount importance for the KMS. Access to keys must be protected physically and logically to avoid any disclosure of critical information, unwanted modifications, unauthorised usage or access. For the physical protection, NIST suggests the adoption of custom hardware solutions, such as hardware security modules. Logical protection measures include encryption, authentication, integrity checks, access control, and accountability.

Physical and logical separation of roles for the actors within the KMS. Access to physical assets, such as, key servers, backup servers, etc. must be limited and monitored. Similarly, from a logical perspective the adoption of different privilege levels can be used to limit the access to critical features of the KMS:

- *Separation of Duties*: no user in the system should have enough privileges to be able to misuse the system. Critical functionalities are split among different members to prevent a single user from having enough information or privileges to maliciously damage the whole system.
- *Least Privilege*: each member or actor of the system is given the least amount of access privileges that allows she/he to perform her/his jobs.

All the above guidelines and principles have been strictly followed and adopted during the design and implementation of the SEkey KMS.

IV. SEkey

In this section we introduce the basic features of SEkey, a KMS that leverages on the features and functionalities provided by the SEcube™ hardware security module. In particular we shall focus on (i) SEkey general architecture, (ii) the concept of *User Groups*, (iii) the different *roles* within the KMS, (iv) how the SEcube™ is profitably employed, (v) the internal structure of SEkey, (vi) the cryptographic keys distribution mechanism, and (vii) the key management feature.

A. SEkey General Architecture

As shown in Figure 2, SEkey manages and distributes cryptographic keys shared among users who are clustered in groups [20], each one being characterised by a custom security policy. The KMS is controlled by an administrator who interacts with the users by means of APIs performing a wide range of actions, such as creating and distributing cryptographic keys, creating and managing users and groups, etc.

A peculiar aspect of SEkey is that each user is forced to make use of a dedicated SEcube™ device, thus implementing a distributed architecture wherein the cryptographic keys are automatically delivered to the users, who securely store them inside their SEcube™ devices. Therefore, the users make use of the KMS together with their SEcube™ devices in order to secure the data they need to exchange or store.

B. User Groups

At the core of SEkey there is the notion of *group* [20], which is the fundamental component used to control the users and the access to the cryptographic keys. Each group consists of an arbitrary number of users and cryptographic keys. Every user of SEkey belongs to a specific set of groups; similarly, each cryptographic key of the KMS is owned by a specific group. A user may belong to several groups, therefore the intersection

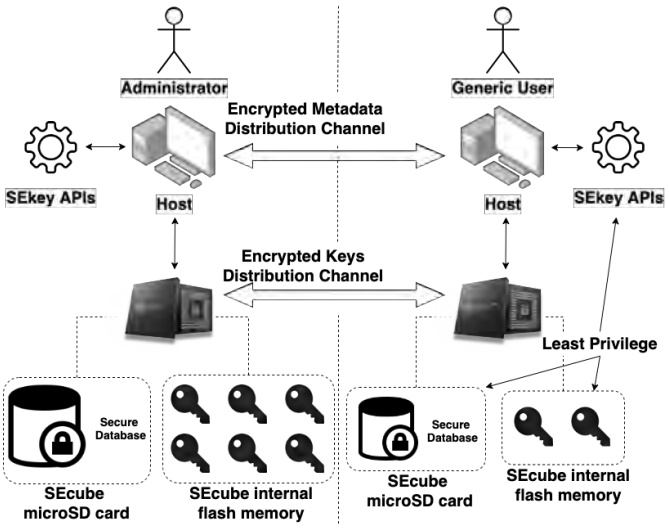


Fig. 2. SEkey General Architecture

of multiple groups may not be empty. On the other hand, the ownership of cryptographic keys is fixed; a key is always owned exclusively by a single group, without any possibility of changing the owner. Notice that the ownership of a key is always referred to a group, never to a single user.

The users gain access only to the cryptographic keys owned by the groups to which they belong; therefore, group members can encrypt shared information using the group cryptographic keys. Two users can share encrypted information only if both belong to at least one group, meaning that they both have access to (at least) one common encryption key. Moreover, each group is associated with a set of *security policies*, detailing specific rules to be followed when managing the security of that specific group. These include, among the others, details about the criptographic algorithms to be adopted, the resource (software, hardware, smart card) to be used for cyphering, the default cryptoperiod of the keys, the schedule of their updating, and so on.

This hierarchy is based on a simple concept: the smaller the group, the higher its security [20]. This idea arises from the assumption that a smaller group involves a reduced number of individuals, therefore the security risks are inherently mitigated because the surface available for a cyber attack is greatly reduced and the sensitive information is shared among a smaller number of people.

C. Roles of the Involved Actors

Actors operating in the SEkey KMS acts either as *administrator* or as *user*. Each role is fixed, meaning that the administrator is not a user and the users cannot act as the administrator.

The *administrator* plays a key role, being the only one having the privilege to modify the configuration of the KMS (i.e., create, distribute, destroy cryptographic keys) and to set up and update groups and users, defining their security perimeters and policies. The SEcube™ device of the administrator contains all the informations managed by the KMS, including all the cryptographic keys; this mainly allows the system to recover from faults that may occurs on the user side. Following the "need-to-know" principle, the administrator shares with the users only the minimum necessary set of information: for example, a user ignores the existence of other people outside of his groups.

Users play a passive role, they can use the KMS but are not allowed to perform any change, neither in the system configuration nor in the involved keys. A user can, in fact, access its own set of cryptographic keys, only; moreover, each key can be used to perform cryptographic operations towards specific recipients only. A user is unable to perform operations which have not been authorized by the administrator (e.g., communicating with users with whom he has got no group in common).

D. SEkey Internal Structure

SEkey, in addition to cryptographic keys, requires to properly manage additional information and metadata which are essential to the system management. To effectively and efficiently tackle this issue, each user of SEkey is given a private instance of the SEcube™ device, which is used to store these critical information items in different locations. In particular, keys are stored in the internal memory of the SEcube™ devices in order to guarantee the highest level of physical protection, whereas the metadata are stored into its MicroSD card. The main reason for this separation is that the size of the internal flash memory of the SEcube™ device is limited to 2 MB, thus it has been reserved to the cryptographic keys.

All the cryptographic primitives are executed by the SEcube™ itself, the user (and administrator as well) only gets the output of those operations, such as encrypted or decrypted data, computed signatures and so on. Moreover, the firmware of the device exposes neither any function to read the content of the internal memory nor the key values in clear, granting a good level of isolation from the main system. In a way similar to the concept of *tokenisations* [11] [12], used in digital payments, where credit card numbers are not sent directly, but instead a mathematically unrelated identifier is shared. Only when the payment has to be processed the unique ID is substituted with the corresponding credit card number; outside of the HSM each key is referred through its own *Unique ID*. It is, hence, impossible to retrieve actual key values because no trace of them can be found anywhere else except the internal memory of the SEcube™ devices.

Since the metadata about keys, users, and groups are stored into a MicroSD card, a different strategy is required to grant a suitable level of security and protection. This alternative strategy relies on SEfile (see Section II): a library of the SEcube™ Open Source SDK that allows to encrypt files and to work with them while keeping everything constantly encrypted on disk. SEfile works together with the open source SQLite[2]

[2]https://www.sqlite.org/index.html

database engine in order to implement a library called 'Secure Database'. In this library, specific for the SEcube™ device, the SQLite database engine has been tweaked to work on a constantly encrypted database while granting confidentiality, integrity and authentication of the DB files thanks to the cryptographic primitives provided by the SEcube™ device.

In addition, SEcube™ is protected by a pair of PIN codes that must be used to access the functionalities provided by the device. Each PIN code is unique to a given SEcube™ and it is associated with a specific privilege level, *admin* and *user*. The PIN codes of each SEcube™ are set during the physical initialization of the device, which takes place before the HSM being physically handed to the user or to the administrator.
The PIN codes of the SEcube™ devices are not related to the actual role of the actors of the KMS. Their only purpose is to stop unauthorised people from accessing the functionalities of the device or limiting the features exposed by the firmware of the SEcube™ to boost the overall security of the system. Following the *Least Privilege* paradigm (see Section III-B), only the minimum amount of information required by each involved actor to perform its operations, is disclosed [15]. For example, each user is provided only with the PIN that grants access to the *user* privilege level of his SEcube™ device while the PIN for the *admin* level is kept secret inside the SEcube™ of the administrator.

E. System Update and Cryptographic Key Distribution

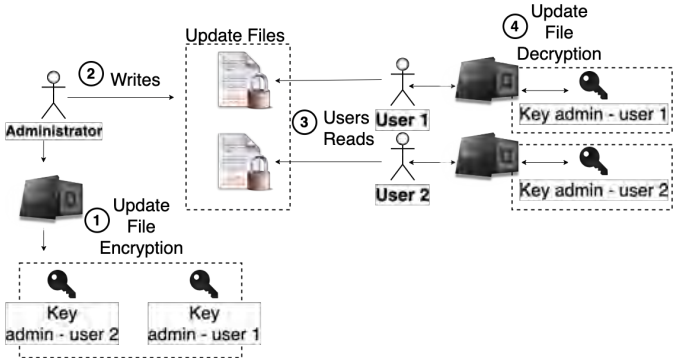


Fig. 3. Update Distribution

Having a distributed architecture where the SEcube™ devices of the users store locally every information that is required for the correct functioning of the KMS, a dedicated secure protocol to share and distribute the data (i.e. groups update, the cryptographic keys and so on) from the administrator to the users is required.

The distribution of the data is always initiated by the administrator, who automatically pushes the data to the users; then the users process these data and store them inside their SEcube™ devices.

This mechanism requires a very simple underlying infrastructure (Figure 3) that relies on update files generated specifically for each user of the system. The update files are encrypted with a key that is known only to the administrator and to the recipient; thus, a secure end-to-end channel is implemented between the host computers of the involved parties. Whenever a new update file is generated by the administrator of SEkey, it is written to a non-volatile memory support that must be accessible also to the users. This non-volatile memory could be anything ranging from a shared disk in a LAN to a cloud service, the only requirement being that all parties involved in the KMS must be able to access to it.
SEkey is configured to automatically generate the update files from the administrator side, and to automatically process them from the user side. The update files contain every data that a given user is entitled to store into her/his personal SEcube™. When SEkey needs to share a cryptographic key from the SEcube™ of the administrator to the SEcube™ of a user, that key must be exported from the HSM of the administrator and written to the update file of the user. The encrypted channel implemented by the update file is not sufficient to protect the key because its value would still be visible to the administrator (the plaintext content of the update file is initially built in the host computer of the administrator, then it is encrypted by the SEcube™ and finally written to the update file). In order to solve this problem, SEkey implements an additional encrypted end-to-end channel, created inside the update file. This channel is built directly between the SEcube™ devices of the involved parties (administrator and user), it allows to export a key from the SEcube™ of the administrator only if that key is already wrapped with another key (which is unique for each user). In this way, the key is already exported outside of the SEcube™ in an encrypted format, guaranteeing that even the administrator cannot see its real value. When the SEcube™ of a user receives a wrapped key, it removes the wrapping and stores the key inside its flash memory, never exposing the real value of the key outside of the HSM.
From a physical point of view, the generation of the cryptographic keys managed by the KMS is always performed inside the SEcube™ of the administrator using a True Random Number Generator embedded in the SEcube™ MCU [16], guaranteeing that each key is random and secure.

F. Key Management Features

The ultimate goal of a KMS is to manage the life cycle of the cryptographic keys. In this sense, each key is characterised by several properties, the most important being the *cryptoperiod* and the *state* (see Section III-B).

The cryptoperiod states how long a key can be used to encrypt data. It can be set, by default, to the value specified by the security policy of the groups that owns the key. However, it could be set to a smaller value when needed; values higher than the default one are not allowed.

The state, instead, determines the current condition of the key. For example, a key can be used to apply cryptographic protection (encrypt data) only if it is in the *active* state; on the other hand, it can be used to decrypt data also if it is not active. Some states, such as *destroyed* and *compromised*, always prevent SEkey from using a key due to security reasons.

Depending on its cryptoperiod and on its state, a key may be eligible for usage. SEkey automatically manages a portion of the life cycle of each cryptographic key, for example it deactivates the keys whose cryptoperiod is expired and it has built-in protection mechanisms to prevent the usage of keys depending on their current state.

When an application needs to perform an encryption operation, it can simply call an API of the KMS that returns the unique identifier of the most secure key to be used, then that identifier is passed to the encryption APIs of the SEcube™. The most secure key to be used in a given situation is determined by the list of the recipients of the data to be encrypted. Here comes into play the concept of *group* (see Section IV-B) so if a user needs to encrypt a message that must be sent to another user, SEkey will automatically search a usable key belonging to the smallest group in common between all the parties involved in the communication, because a smaller group is considered to be safer. The same holds if a user wants to encrypt data for private usage, for example before storing them on a cloud server. In that case the user will specify himself as the only recipient, so SEkey will search for a usable key belonging to a group where that user is the only member.

In addition to the keys managed by the KMS, additional cryptographic keys are required to properly manage the system. These keys are not under the direct control of the KMS or the administrator, but are generated automatically by the system. are not visible to the user, and are used to encrypt data locally to each SEcube™. For example, every SEcube™ generates a unique key that is used to encrypt the metadata database of SEkey.

V. Conclusions

In this paper, SEkey was presented, a key management system that leverages the peculiar features and functionalities of the SEcube™ hardware security module to provide all what is required to securely manage cryptographic keys.

During the design of SEkey, all the most important security dictates provided in the NIST guidelines were followed. Each key is associated with a cryptoperiod and a state. Seven different states are used to determine the type of operations that a key can perform. Moreover, following the 'Least Privilege' principle two actors with different privileges have been identified in the KMS, the *administrator* and the *user*: the former having the full privilege to perform any modification to the KMS data while the latter can just use the KMS passively without any authority to make changes.

The SEkey KMS is based on a distributed structure and its users are organised according to a particular hierarchy that provides multiple groups, each characterised by specific security policies. Users can communicate and share information with each other by means of symmetric cryptographic keys shared within the group. Each actor in the KMS has its own SEcube™ HSM and all the cryptographic keys and critical items are stored securely in the internal device flash memory. Moreover, all the cryptographic primitives are provided by the SEcube™ itself, hence keys never leave the device when performing crypto operations and are never exposed in clear. The keys that are distributed by the administrator are over-encrypted with a unique key shared only between the administrator and the user who must receive them. To limit the use of the device internal memory, all the metadata handled by the KMS are saved, on a MicroSD card connected to the SEcube™, in an always encrypted database, thus guaranteeing the integrity, confidentiality and authenticity of these data. Since the internal memory of SEcube™ is limited to 2MB available, the adopted approach allows storing inside a single flash memory sector (128KB) up to 4096 different keys (assuming a key size of 256 bits).

As far as future improvements there are the following aspects are going to be tackled in the near future:

- Management of session keys: keys that can be generated, used and dismissed within a group when there is the need of instantiating a communication channel. In this way it is possible to better separate keys that can be used to cryptographically secure data at rest (e.g., files) and data in motion (e.g., calls). Groups can internally manage the creation of these type of keys, using for example a contributory key agreement protocol, without querying the central manager.
- Improvement in the internal flash memory management of the device: since flash memories have a limited amount of write operations that can be performed, having to replace every now and then keys inside it can quickly wear out memory.
- Implementation of a PUF inside the device: this can be used either as a strong private cryptographic key, used for example for the metadata database encryption, or as a unique key shared by the administrator and each user used for the encryption of SEkey update messages.
- Addressing the problem of *non-repudiation* in group encryption. Methodologies exists involving either asymmetric encryption, such as *Ring Signature* [21] or *Threshold Signature* [1], or symmetric encryption such as the use of trusted third party top provide a One Time Password to be used in the signing process [17].

VI. Acknowledgments

The activities presented in the present paper are partially supported by the *European Union's Horizon 2020 research and innovation programme*, under grant agreement No. 830892, project SPARTA and by *B5 Labs Ltd.*.

References

[1] Michel Abdalla, Sara Miner, and Chanathip Namprempre. Forward-secure threshold signature schemes. In *Cryptographers' Track at the RSA Conference*, pages 441–456. Springer, 2001.

[2] Y. Amir, Y. Kim, C. Nita-Rotaru, J. L. Schultz, J. Stanton, and G. Tsudik. Secure group communication using robust contributory key agreement. *IEEE Transactions on Parallel and Distributed Systems*, 15(5):468–480, 2004.

[3] E. Barker. Recommendation for key management: Part 1 - general. *NIST, Tech. Rep*, 2020.

[4] W. C. Barker. Guideline for identifying an information system as a national security system. *NIST, Tech. Rep*, 2003.

[5] M. Bollo, A. Carelli, S. Di Carlo, and P. Prinetto. Side-channel analysis of secube™ platform. In *2017 IEEE East-West Design Test Symposium (EWDTS)*, pages 1–5, 2017.

[6] CRYPTOMAThIC. Selecting The Right Key Management System. https://www.cryptomathic.com/hubfs/Documents/White_Papers/Cryptomathic_White_Paper_-_Selecting_The_Right_Key_Management_System.pdf, 2019. [Online; accessed 22-July-2020].

[7] K. Dempsey, M. Nieles, and V. Y. Pillitteri. An introduction to information security. *NIST, Tech. Rep*, 2017.

[8] M. Fornero, N. Maunero, P. Prinetto, G. Roascio, and A. Varriale. SEcube Open Security Platform - Introduction. https://www.secube.eu/site/assets/files/1218/wiki.pdf, 2019. [Online; accessed 22-July-2020].

[9] M. Fornero, N. Maunero, P. Prinetto, G. Roascio, and A. Varriale. SEfile Documentation. https://www.secube.eu/site/assets/files/1218/wiki.pdf, 2020. [Online; accessed 22-July-2020].

[10] V. Gopal, S. Fadnavis, and J. Coffman. Low-cost distributed key management. In *2018 IEEE World Congress on Services (SERVICES)*, pages 57–58, 2018.

[11] Gabriel Babatunde Iwasokun, Taiwo Gabriel Omomule, and Raphael Olufemi Akinyede. Encryption and tokenization-based system for credit card information security. *International Journal of Cyber Security and Digital Forensics*, 7(3):283–293, 2018.

[12] Ronald Julien Jr. *The cybersecurity aspects of Apple Pay*. PhD thesis, Utica College, 2016.

[13] T. Mandt, M. Solnik, and D. Wang. Demystifying the secure enclave processor. *Black Hat Las Vegas*, 2016.

[14] Sandro Rafaeli and David Hutchison. A survey of key management for secure group communication. *ACM Computing Surveys (CSUR)*, 35(3):309–329, 2003.

[15] F. B. Schneider. Least privilege and more [computer security]. *IEEE Security Privacy*, 1(5):55–59, 2003.

[16] STMicroelectronics. AN4230 Application Note - STM32 microcontroller random number generation validation using the NIST statistical test suite. https://www.st.com/resource/en/application_note/dm00073853-stm32-microcontroller-random-number-generation-validation-using-the-nist-statistical-test-suite-stmicroelectronics.pdf, 2020. [Online; accessed 22-July-2020].

[17] International Telecommunication Union. X.1156: Non-repudiation framework based on a one-time password. https://www.itu.int/rec/T-REC-X.1156-201306-I/en, 2014. [Online; accessed 22-July-2020].

[18] A. Varriale, E. I. Vatajelu, G. Di Natale, P. Prinetto, P. Trotta, and T. Margaria. Secube™: An open-source security platform in a single soc. In *2016 International Conference on Design and Technology of Integrated Systems in Nanoscale Era (DTIS)*, pages 1–6, 2016.

[19] Antonio Varriale, Giorgio Di Natale, Paolo Prinetto, Bernhard Steffen, and Tiziana Margaria. Secube (tm): an open security platform-general approach and strategies. In *Proceedings of the International Conference on Security and Management (SAM)*, page 131. The Steering Committee of The World Congress in Computer Science, Computer …, 2016.

[20] Antonio Varriale, Paolo Prinetto, Alberto Carelli, and Pascal Trotta. Secube (tm): Data at rest and data in motion protection. In *Proceedings of the International Conference on Security and Management (SAM)*, page 138. The Steering Committee of The World Congress in Computer Science, Computer …, 2016.

[21] Fangguo Zhang and Kwangjo Kim. Id-based blind signature and ring signature from pairings. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 533–547. Springer, 2002.

Hardware-based Capture-The-Flag Challenges

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

Gianluca ROASCIO
Politecnico di Torino
CINI Cybersecurity National Lab.
Turin, Italy
gianluca.roascio@polito.it

Antonio VARRIALE
Blu5 Labs Ltd.
Ta' Xbiex, Malta
av@blu5labs.eu

***Abstract*—In a world where cybersecurity is becoming increasingly important and where the lack of workforce is estimated in terms of millions of people, gamification is getting a more and more significant role in leading to excellent results in terms of both training and recruitment.**

Within cybersecurity gamification, the so-called Capture-The-Flag (CTF) challenges are definitely the corner stones, as proved by the high number of events, competitions, and training courses that rely on them. In these events, the participants are confronted directly with games and riddles related to practical problems of hacking, cyber-attack, and cyber-defense.

Although hardware security and hardware-based security already play a key role in the cybersecurity arena, in the worldwide panorama of CTF events hardware-based challenges are unfortunately still very marginal.

In the present paper, we focus on hardware-based challenges, providing first a formal definition and then proposing, for the first time, a comprehensive taxonomy. We eventually share experiences gathered in preparing and delivering several hardware-based challenges in significant events and training courses that involved hundreds of attendees.

***Index Terms*—cybersecurity, education, gamification, capture-the-flag, challenges, hardware, hardware security.**

I. Introduction

In recent times, the world is experiencing a digital revolution that leaves no aspect of our life uncovered. Our job, the management and purchase of good and services, and even the organisation of our free time inevitably rely on constantly-connected digital devices. As a consequence, the issue of data security can no longer be ignored and it must be addressed at every level, from the awareness rising of any citizen, up to the massive investments by public and private sectors in order to increase the number of available experts in cybersecurity. Today, we see a very strong push towards hiring Research and Development specialists in cybersecurity, in a plenty of fields and domains. The number of projected unfilled jobs worldwide in cybersecurity has recently been estimated in 3.5 M by 2021 [18]. Such a huge number definitely requires a significant push by institutional education centers such as schools and universities. Nevertheless, even if the number of dedicated academic courses, BS and MS curricula is growing, they still lack a *practical imprint*, on which the attackers are instead very well prepared [27]. Without a significant effort in this direction, presenting cybersecurity from theory-oriented point of view is likely to appear as boring and therefore not attractive to many students.

An important different paradigm in cybersecurity teaching is the one heavily exploiting *gamification* [30] [35]: students are asked to directly face security problems by solving riddles and challenges related to the breakdown or the decryption of software, communication systems or devices, or by implementing countermeasures to prevent attacks by opposing teams. Within cybersecurity gamification, the so-called *Capture-The-Flag (CTF) challenges* are definitely the corner stones, as proved by the high number of events, competitions, and training courses that rely on them. In these events, the participants are confronted directly with games and riddles related to practical problems of hacking, cyber-attack, and cyber-defense. CTF challenges are the basis of the so-called *Capture-The-Flag (CTF) competitions*, where the aim is in fact to extract from the challenge a unique string, the *flag*, which certifies the success.

The results of the CTF in terms of education and creation of new practical knowledge have been recognised by various studies [29] [43] [39], also in relation to their adoption in the context of university courses on computer security [34]. CTF competitions are therefore a valid approach to try to fill the workforce gap, mainly thanks to their attractive power among new generations and to their ability to make participants develop their *adversarial thinking*, which has proved to be essential in defending infrastructures from malicious cyber-attacks [42].

On the basis of these considerations, the Italian CINI Cybersecurity National Laboratory[1] has been developing the *CyberChallenge.IT*[2] program since 2017. CyberChallenge.IT is, in fact, the main Italian initiative aiming at identifying, attracting, recruiting, and placing the next generation of IT security professionals, thus seeking to reduce the lack of IT workforce at the national level. Its target are young talents (aged 16-23) and the 2020 edition has involved more than 4,400 of the best students who live and study in Italy. To create and grow such a community of young cyber-defenders, the program offers training opportunities to stimulate interest in STEM disciplines and, in particular, in information and computer security. Participants also have the opportunity to get in direct contact with IT companies working in the field, which actively contribute to their orientation and professional training. The program combines traditional training activities

[1] https://cybersecnatlab.it/
[2] https://www.cyberchallenge.it/

with a gamification-oriented approach which requires the participation in on-line competitions where different scenarios of networks and real work environments are simulated. The model is unique on the international scene; in fact, not only it exploits gaming as an instrument for attracting young people, but it also offers a multidisciplinary training.

The training process ends with two final competitions, organized at the training node level and at the national level, respectively. The former is a Jeopardy style CTF (see Section II) run concurrently by all the attendees of all the training nodes. The latter, which is in fact the Italian CTF championship in Cybersecurity, is an Attack-Defense style CTF (see Section II), attended by teams of 6 members each, one per training node, and planned each year in a specific location. In 2020, both the competitions have been organized remotely, due to Covid-19 restrictions. Both the competitions exploit infrastructures in terms of servers and software applications, completely developed in-house and managed by the Cybersecurity National Laboratory. Similarly, all the challenges used in both the Jeopardy and the Attack-Defense competitions are brand new and developed in-house. Since the 2020 edition, students experience hardware security techniques and then *hardware-based challenges* during the final competition.

CTFs usually focus on "mainstream" aspects of cybersecurity, such as challenges based on web exploits in which, for example, it is possible to exploit SQL injection [16] or Cross-Site Scripting (XSS) [15] to retrieve the flag, or a step-based challenge where the interaction with a command line is offered by a vulnerable system that hides the flag, for example, in the home folder of some user whose login needs to be cracked, or in some software to be attacked through code injection [17]. This prevalence is caused by the fact that these issues are very popular, and related problems with possible defense techniques have been studied more in depth and for a longer time. However, it is to be pointed out that all layers of an IT system can be subject to threats, from the highest application layers down to the hardware level. Hardware components are subject to intrinsic vulnerabilities and are exposed to particular attacks [45], which can determine an even more marked danger. In fact, hardware is not patchable as a piece of software, and if present, the vulnerability remains until the component is active. Furthermore, hardware is *the root* of systems: if hardware is compromised, all upper layers could be compromised as well, even if protected against web or software attacks [26]. The theme should not be underestimated or downgraded to a *niche* theme, inaccessible to most: it should instead be included in the ecosystem of cybersecurity education and training, included CTF competitions as well.

The present paper focuses on hardware-based challenges, providing first a formal definition and then proposing, for the first time, a comprehensive taxonomy. Some examples of real challenges are then presented, each classified according to the proposed taxonomy. We eventually share experiences gathered in preparing and delivering several hardware-based challenges in different environments. The sequel of the paper is organized as follows. Section II provides a general background on CTF competitions; Section III details hardware-based challenges and proposes a new taxonomy, as well as an overview of the hardware-based challenges proposed in some famous CTF competitions around the world; Section IV reports some examples of hardware-based challenges implemented this year during the CyberChallenge.IT event; Section V concludes the paper.

II. Background on CTF Competitions

A Capture-The-Flag challenge is a game in which the goal is breaching into one or more vulnerable IT assets (websites, files, databases, network devices, hardware devices, and so on) to guess or get a *flag* [32]. The flag is a unique string, decided by the organizers and formatted in a competition-specific manner, which certifies the success in the challenge. Individuals or teams participating to CTF competitions get points for each correct flag submitted to the competition organisers, and the winner is usually the individual or team owning the highest number of points at the end of a given predefined time slot.

Three main different CTF challenge types exits, based on execution modalities and involved actors:

- *Jeopardy*: Participants are asked to face a vulnerable system which hides the flag. The flag can be "captured" by exploiting the vulnerabilities that have been artificially inserted into the system by the competition organizers. Participants may be grouped in teams, but there is no interaction among the teams. The only opponent is the challenge itself.
- *Attack/Defense*: Participants are grouped in teams and each team is given an instance of a system injected with several vulnerabilities. All the instances get the same vulnerabilities and are connected to a same network. The competition includes two phases: for a first period of time (e.g., one hour), each team can access its instance, only, and, during this slot, the team should identify and fix the vulnerabilities on its own instance. In this way they can prevent other teams from capturing their flag exploiting these vulnerabilities during the next phase. In a second phase, connection is opened and each team is free to access the instances of the opponent teams and capturing their flags if the vulnerabilities present in their instances have not been properly patched during the first phase. Points are awarded based on three factors: (i) the number of flags captured on the instances of other teams (*attack points*), (ii) the number of flags stolen by other teams from your instance (*defense points*), (iii) the percentage of time the services remain up and work properly (*SLA points*). With respect to Jeopardy-style, these challenges allow participants to gain experience on both offensive and defensive skills.
- *King of the Hill*: It is a slight variant of Attack/Defense CTF, in which participants are usually grouped in teams, and the goal is taking and holding control of a machine or a network. The challenge last for a given period of time,

and at the end, the team that held the system longest is the winner.

CTFs are usually clustered according to the topics they deal with, as:

- *Binary*: This category includes all those challenges that require the exploitation of a vulnerable software application. The name stems from the abstraction level exploited during the attack, i.e., the machine binary code, often resorting to disassembly and debugging tools. This class can be further split into:
  - *Reversing*: The challenges are based on the backwards reconstruction of the behavior of the application, in order to allow, for instance, a particular interaction to retrieve the flag. Beyond the knowledge of programming languages, a good familiarity with static code analysis tools such as decompilers and disassemblers is often required.
  - *Pwn*: These are the challenges that most closely resemble hackers' activities in the collective imagination, e.g., breaking a remote vulnerable service on a server. The exploit can be carried out by injecting binary instructions into the application's memory through a breach opened by a vulnerability, or by hijacking the execution towards blocks of hidden code or spread bytes that were not originally intended to be executed in that order. For these challenges, the vulnerable binaries can be presented either as *white-box* if source files are made available, or *black-box* if no file is attached.
- *Web*: This category includes all the challenges dealing with vulnerable web services, susceptible to attacks based on command or code injections, which allow to retrieve information that is originally not accessible, including the flag. Examples include challenges based on web login crack, malicious SQL query injections, tampering with cookies, etc.
- *Crypto*: The challenges consist in breaking an encryption scheme to decipher a message that directly or indirectly contains the flag. The encryption scheme can be either a classic one but implemented in a vulnerable way, or a brand new one to be reversed. These are usually the longest challenges in terms of time, because they may require an automatic breaking phase due, for example, to the execution of an *ad hoc* script written by participants. Mathematical knowledge of combinatorics, prime numbers, modular arithmetic are usually very helpful.
- *Forensics*: This category of challenges takes its name from the fact that the techniques used to capture the flag mimic the typical forensic approaches adopted by law enforcement and investigation agencies. They very often exploit steganography, including, for instance, malformed files, packet captures, .jpg or .png files modified to hide texts or executable pieces of code. By digging into these files with scripts and tools, participants can extract data (that are often encrypted) to recover the flag.
- *Networking*: In these challenges actions typical of the network domain (such as: breaking firewalls, deceiving access policies, attempting spoofing attacks and poisoning of network protocols, or reconstructing a message from individual packets) must be exploited to capture the flag.
- *Miscellaneous*: The challenges typically span several non-technical topics and their resolution usually requires just basic logic and/or reasoning efforts, thus to make them beginner-friendly.
- *Hardware*: These challenges will be extensively discussed in the next Section.

III. Hardware CTF

A. A Look Around

As already mentioned, in the context of the CTF competitions organized around the world, the topic of hardware security today still plays a very marginal role, when not present at all. This is mainly due to the relative novelty of the topic, which has begun to spread only in recent years. Hence follows a low amount of specific skills, compared with the much greater amount of experts in the fields of cryptography, reversing, software security, among the organizers of the competitions as well. Therefore, a state-of-the-art of the topic limits to an overview of the few events that worldwide include hardware-related challenges.

The *Hardwear.io* platform [23], which includes hardware security researchers from all over the world and organizes courses, conferences and webinars on the topic, hosts a hardware-oriented CTF competition since 2017. The proposed challenges typically cover various themes, such as RFID, Bluetooth, automotive components, side-channel analysis, (de)soldering and radio. Proper sets of physical tools needed for the challenges and a guidance on how to use them are usually provided to the participants.

Riscure [25], an important security evaluation laboratory specialized in embedded systems and IoT security, organized RHme (Riscure Hack me) from 2015 to 2018: a CTF event mainly oriented to safety in the automotive environment and based on the use of Arduino™ products [10] for the implementation of the challenges [3] [4] [8]. The LiveOverflow channel maintains a collection of videos regarding the challenges of the event and their solutions [5]. Although very innovative and well implemented, many of these challenges are based on attacks on cryptographic protocols (e.g., a length-extension attacks to a SHA implementation) or communication protocols (e.g., UART), which do not require any specific knowledge of the hardware domain. In these challenges, the physical boards just play the role of a mere support for the execution of the challenge, in a manner no different from that of PCs, servers, virtual machines, switches and all the other devices used in challenges of any type. As we shall point out in the sequel of paper, we do not consider the above challenges as "true" hardware-based CTF challenges.

The *Hack@DAC* [21] hardware security contest has been held within the Design Automation Conference (DAC) [19]

since 2017. It is a competition focused on the topic of microarchitectural and side-channel flaws in chips [33] [41] [36]. Participating teams (students and industrial teams as well) are given a design of a vulnerable chip to be studied before the competition. The aim is to identify the greatest number of security problems. The winners of this first phase then participate in the CTF competition held live at the conference: here, the teams are assigned a new design of a vulnerable SoC, and must take advantage of their previous experience to find as many vulnerabilities as possible in a given time slot. At the end, the winner is the team that has submitted, in the format of flags, the greatest number of problems in the design.

Some general CTF events tried to incorporate hardware-based challenges into their programs. *Chujowy CTF* [11] introduced challenges aimed at finding vulnerabilities within the Verilog code of an automotive processor based on RISC-V [13] [12]. The *Google Capture The Flag* event [20] introduced some hardware-oriented challenges as well. In the 2017 edition, a challenge which consisted in cracking a slot machine, required to physically connect to the pins of the Arduino™ board which controlled the machine in order to extract the flag [6]. Other challenges that required to reverse HDL code or schematic hardware components were included in the 2018, 2019, and 2020 editions [7] [9] [14].

B. Definition and Taxonomy

The purpose of this subsection is twofold: we first provide a definition of *hardware-based CTF challenge* and then propose a brand-new taxonomy of hardware-based challenges. At the authors' best knowledge, both the definition and the dimensions of the taxonomy are introduced here for the first time and they both stem from the experiences authors collected while preparing and delivering several hardware-based challenges in significant competitions, talent scouting programs, training activities, and BS and MS level courses that globally involved hundreds of attendees.

A *hardware-based CTF challenge* is a challenge in which the challenger must exploit her/his knowledge about digital hardware (including methodologies and technologies related to design, validation, verification, testing, maintenance, etc., at all the abstraction levels), in order to capture a flag consisting in identifying, remediating, or exploiting vulnerabilities [45] artificially introduced either in the design or in the actual implementation of the hardware structure of a digital system.

The above definition has some relevant practical implications, among which we would like to point out the following ones:

1) The fact that a challenge simply "rely" on a hardware device does not imply that the challenge is a hardware-based CTF challenge. At the ultimate end, each program runs on hardware, so "running on a hardware device" cannot be a sufficient condition;
2) In a true hardware-based CTF challenge, capturing the flag must require a significant knowledge about digital hardware and cannot be successfully solved exploiting just other lateral knowledge;
3) In a true hardware-based CTF challenge, the flag could not be captured by just exploiting some vulnerabilities artificially introduced either in the software applications that runs on the hardware device, or in the communication or security protocols that are adopted by that device;
4) A true hardware-based CTF challenge can be implemented and proposed in real competitions without resorting to any "physical" hardware device, since it can rely on some particular features or aspects of the design of the device, which can be provided to participant via description files or proper EDA environments and tools.

Starting from the above definition, let's now propose a new taxonomy for hardware-based CTF challenges. It relies on five orthogonal dimensions: (i) the challenge purpose, (ii) its difficulty, (iii) its execution mode, (iv) its topic, and (v) the hardware device description. Let's analyse each dimension in details:

1) **Challenge Purpose**: challenges have to be planned differently according to their ultimate purpose, distinguishing among:
   - *Training*: in this case the challenge should be organized in such a way to smoothly drive the students through the different learning steps, usually characterised by an increasing complexity;
   - *Competition in Jeopardy style*: (see Section II): these challenges are usually more complex and hard-to-solve versions of the challenges used for training, where additional tricks are intentionally inserted according to the difficulty level of the competition;
   - *Competition in Attack/Defense style*: (see Section II): these are definitely the most complex hardware-based challenges, since they pose a lot of severe and hard constraints, including, among the others, the fact that any team must get a copy of the instance and that, during the second phase of the competition, each instance can be concurrently accessed by all the teams and by the game server. This practically means that, when a physical hardware device is involved, each instance must be equipped by a custom (software) wrapper in order to properly queue, manage, and serve all the incoming concurrent requests.
2) **Challenge Difficulty**: each challenge should be characterized by a proper ranking of its difficulty. To our best knowledge, unfortunately no consolidated official ranking schemes exist today. Consequently, it is usually up to the challenge's authors to provide a reasonable ranking, based on their experiences in training and gaming. We usually adopt a 5-value ranking, 1 being the easiest and 5 the hardest.
3) **Challenge Execution Mode**: it mainly deals with the tools provided to the attendees to solve the challenge. Three possibilities are usually exploited:
   - *By-hand*: Participants are given a design representation of a digital hardware (usually HDL code

or schematics), and the challenge can be solved manually just carefully analyzing the provided design description, without the need to resort to any specific tool. This kind of challenges are definitely the easiest and cheapest to implement and they just require expertise and knowledge in hardware design and test to be solved.
- *EDA-tool-based*: To solve the challenge, the participants have to resort to the facilities offered by a specific EDA platform (typically simulators and/or automated synthesis tools), made (fully or partially) available during the competition. Participants can exploit the provided tool to access a "model" of the hardware device, in which the vulnerabilities have been inserted. Note that, in this case, an instance of the selected platform must be made available to each participant and, in some cases, custom "wrappers" have to be designed in order to prevent participants from using the whole set of capabilities offered by the platform, since they could exploit some of these facilities to find a fastest and trivial ways to capture the flag.
- *Hardware-device-based*: In this case each participant must face a real hardware device (typically a small system, a PCB, or a development kit) that somewhere and somehow stores the flag to be captured. In some cases, an FPGA-based implementation/emulation of the target hardware device can be profitably exploited. Note that this case poses some severe issues in term of scalability, since during the competition each participant (or team of participants) must be given a different instance of the hardware device, regardless the competition type. Practically, it can be effectively adopted in the training phases and in teaching courses, where the hardware resources can be effectively shared in time among the attendees.

It is worth mentioning that, from a conceptual point of view, a fourth alternative, completely based on a *pure software emulation* of the hardware device could theoretically be adopted. In our experience, such a solution is mostly ineffective, due to practical difficulties in completely emulating via software, at the same time: (i) the expected hardware behavior, (ii) the set of vulnerabilities to be inserted, and (iii) their possible remediations.

4) **Challenge Topic**: several topics can be covered, including, among the others, the following ones:
   - *Hardware Trojans*: Participants are provided with a digital hardware into which a hardware trojan [46] has been artificially inserted. The identification and/or the exploitation of the trojan lead the participants to capture the flag.
   - *Unprotected test infrastructures*: In these challenges, the flag can be obtained by a clever exploitation of a test infrastructure available in the hardware device. These can range from the IEEE standard 1500 - Standard for Embedded Core Test [1] to the 1149.1-2013 - IEEE Standard for Test Access Port and Boundary-Scan Architecture [2] and to simple scan chains [28], all left accessible.
   - *Undocumented functions and features*: In additions to the hardware descriptions or implementations, participants are provided with a related documentation in which some peculiar features are deliberately omitted. These may include, for instance, machine instructions, components or undocumented side effects of the joint use of multiple documented components [31]. The flag can be captured only by exploiting (one of) these hidden features.
   - *Design bugs and flaws*: The hardware that participants have to deal with includes some design bugs or flaws [45] that introduce a vulnerability, through which the flag can be reached. Examples include, among the others, incorrectly-implemented machine instructions, internal race condition for which sensitive information can be released, etc.
   - *Side-channel Attacks*: Participants are given a hardware device vulnerable by side-channel attacks [40], such as timing or power attacks [38] [37]. Participants must be equipped with a set of tools that allow them to perform the attack and capture the flag.
   - *Weak implementations of hardware-based security modules*: The hardware delivered to the participants belongs to one of the families of modules used for security (e.g., hardware ciphers, random number generators, authenticators, etc.), but that has been designed and implemented introducing some weakness or vulnerabilities that can be exploited to get to the flag.

5) **Hardware Device Description**: when a description of the hardware design has to be provided, two orthogonal additional dimensions have to be considered, the *Abstraction Level* and the *Representation Domain*:
   - **Abstraction Level**: it identifies the level of details provided in the system description: it typically ranges from *System* to *Register-Transfer (RT)* to *Logic* level. Very seldom lowest abstraction levels are used.
   - **Representation Domain**: the provided descriptions can belong to the *Behavioral* or *Structural* domain. In the former case, the *behavior* of the target hardware system is provided in terms of properties (both functional and non-functional ones), which define what the system does and the circumstance under which it operates; in the latter case the *structure*, (i.e., the topology) of the target system is provided in terms of a set of functional building blocks, properly interconnected.

In conclusion, it is worth pointing out that hardware-based challenges involving *invasive attacks* [45] are usually not implemented, since they require the availability of advanced (and often expensive) tools and equipment that, in turns, require a very high degree of expertise to be properly and safely used and exploited.

IV. Our Experience

In this Section, we briefly introduce some hardware-based challenges that we prepared and delivered in different environments, including, among the others: (i) the training of TeamItaly (the Italian Team of cyber-defender that got the silver medal at the last European Championship in Bucharest, on November 2019), (ii) significant competitions, (iii) CTF training courses that involved hundreds of attendees with different backgrounds, and (iv) University courses in Cybersecurity. In particular, we shall focus on 4 different challenges. Readers interested in additional technical details are kindly invited to directly contact the paper's authors.

A. TIGER21X

- *Challenge Purpose*: Training
- *Challenge Difficulty*: 5/5 (hard)
- *Challenge Execution Mode*: By-hand
- *Challenge Topic*: Undocumented functions and features
- *Hardware Device Description*: RT-level structural.

Challenge Description: participants are faced with a simple custom processor, implemented as a reduced variant of the RISC DLX ISA [44]. In particular, they are provided with (i) RT-level structural description of the device, (ii) VHDL behavioral description of its control unit and (iii) device technical documentation. The processor implements an undocumented machine-level instruction, and namely an *indirect jump*, i.e., a jump instruction whose destination address is stored into one of the user-accessible general-purpose register. The flag to be captured is the label of the undocumented machine instruction.

In order to successfully solve the challenge, participants have first to find the proper mapping between documented machine instructions and their actual implementation, by reversing the control bits that the execution of each instruction activates in the data path. Of course, opcodes present in the control unit VHDL code have been labeled with non-speaking names not to make trivial the mapping. The additional undocumented instruction was properly hidden among the others, and to identify it participants must understand the meaning of each control signal issued by the control unit and to note the strange behavior of the processor when the undocumented machine instruction is executed.

B. AUTH_98X276YC

- *Challenge Purpose*: Jeopardy Competition
- *Challenge Difficulty*: 4/5 (medium-hard)
- *Challenge Execution Mode*: By-hand
- *Challenge Topic*: Hardware Trojan
- *Hardware Device Description*: Logic-level structural.

Challenge Description: participants are faced with a hardware device implementing an access enabler which grants access when a user-provided key matches the one previously stored inside the device, by asserting a PASS_FAIL output signal. The device includes a hardware trojan which, when activated, serially outputs the content of the stored key. Participants are given a file containing the behavioral specifications of the circuit and its *netlist*, i.e., a structural description at the logic abstraction level. The flag to be captured is the sequence of values to be assigned to input signals to get the key stored inside the device. The solution can be reached first by noticing that the internal registers are implemented in such a way that they could behave as shift registers, and then identifying the trojan activations sequence; this requires a detailed analysis of the provided netlist.

C. BROK_11491

- *Challenge Purpose*: Jeopardy Competition
- *Challenge Difficulty*: 3/5 (medium)
- *Challenge Execution Mode*: EDA-tool-based
- *Challenge Topic*: Unprotected test infrastructures
- *Hardware Device Description*: RT-level structural.

Challenge Description: participants are asked to capture a flag consisting in the value stored into the Device Identification Register of a simple hardware device designed to be compliant with the IEEE 1149.1 standard. To get it, they are given: (i) the device data sheet, which includes all the details about the TAP implementation, (ii) the RT-level structural VHDL description of the device, (iii) the possibility of using a simulator, properly wrapped in order to allow the users just to force the device's primary inputs, to read its primary outputs, and to run simulation campaigns.

A variation of the challenge can be proposed, in which the simulator is replaced by an actual implementation of the hardware device, typically resorting to a FPGA. In this case, an additional environment that allows participants to interact with the device must be provided.

D. CrashCube

- *Challenge Purpose*: Jeopardy Competition
- *Challenge Difficulty*: 3/5 (medium)
- *Challenge Execution Mode*: EDA-tool-based
- *Challenge Topic*: Weak implementations of hardware-based security modules
- *Hardware Device Description*: The physical device is provided, along with System-level documentation.

Challenge Description: participants are asked to investigate how extracting sensible data and keys from the secure flash embedded in the USB cryptographic token emulated by the SEcube™ development kit [24]. The secrets can be extracted exploiting either a hardware (semi-permanent) vulnerability of the chip, based on the IEEE 1149.1 standard, or a firmware vulnerability based on a mismanagement of the internal flash segments. To extract the secrets, they are given: (i) device data sheet, including the JTAG semi-permanent and permanent

states' description, (ii) partial info on the serial communication protocol, (iii) JTAG programmer and flash programming tool.

A variation of the challenge can be proposed, in which the SEcube™ exposes its internal bus between the embedded application secure processor and the embedded smart card. In this case, participants may benefit from extra information retrieved through a probe on the exposed bus, as it happened in many real cases where the CPU and the smart card are separate devices mounted on a PCB (e.g., Ledger cryptocurrency hardware wallet [22]).

V. Conclusions

In this paper, leveraging the experiences we gathered form different training opportunities and official competitions, we defined the concept of *hardware-based CTF challenge*, i.e., a challenge based on hardware-related security issues. In addition, for the first time, we proposed a taxonomy and presented some challenges we adopted in different situations.

Although hardware security is getting increasing interest within the cybersecurity community, its role in international competition is still marginal. In Italy, the 2020 edition of the *CyberChallenge.IT*[3] program for the first time included a complete week devoted to hardware security and in the final national competition, in Jeopardy style, attended by 400+ participants, we proposed a set of three hardware-based CTF challenges.

A plenty of work still needs to be done, especially in the direction of defining some shared and agreed metadata for the challenges, on their classifications, and on their sharing. Additional issues concern identifying and experiencing some viable solutions for properly and effectively including hardware-based CTF challenges within different training environments, including, among the other, from the one hand, professional hybrid Cyber Ranges and, from the other, university BS and MS courses.

Authors are interested and available to share experiences on the various issues outlined in the present paper.

VI. Acknowledgments

The activities presented in the present paper are partially supported by: (i) the European Union's Horizon 2020 research and innovation programme, under grant agreement No. 830892, project SPARTA, (ii) the Italian CINI Cybersecurity National Lab. via the program *CyberChallenge.IT*, and (iii) Blu5 Labs in Malta.

References

[1] IEEE 1500 Standard for Embedded Core Test (SECT). http://grouper.ieee.org/groups/1500/index.html, 2005. [Online; accessed 03-August-2020].

[2] 1149.1-2013 - IEEE Standard for Test Access Port and Boundary-Scan Architecture. https://standards.ieee.org/standard/1149_1-2013.html, 2013. [Online; accessed 24-July-2020].

[3] GitHub - Riscure/RHme-2015: RHme+ 2015 challenge. https://github.com/Riscure/RHme-2015, 2016. [Online; accessed 23-July-2020].

[4] GitHub - Riscure/RHme-2016: RHme2 challenge (2016). https://github.com/Riscure/RHme-2016, 2017. [Online; accessed 23-July-2020].

[5] Riscure Embedded Hardware CTF - LiveOverflow. https://old.liveoverflow.com/rhme/index.html, 2017. [Online; accessed 23-July-2020].

[6] slot machine write-up (Google CTF 2017 Finals). https://blog.bushwhackers.ru/slot-machine-write-up-google-ctf-2017-finals/, 2017. [Online; accessed 24-July-2020].

[7] 2018-06-23-Google-CTF-Quals - 220 Misc / Wired CSV. https://github.com/EmpireCTF/empirectf/blob/master/writeups/2018-06-23-Google-CTF-Quals/README.md#220-misc--wired-csv, 2018. [Online; accessed 24-July-2020].

[8] GitHub - Riscure/RHme-2017: Riscure Hack Me embedded hardware CTF 2017-2018. https://github.com/Riscure/RHme-2017, 2018. [Online; accessed 23-July-2020].

[9] CTFTime.org / Google Capture The Flag 2019 (Quals) / flagrom / Writeup. https://ctftime.org/writeup/15870, 2019. [Online; accessed 24-July-2020].

[10] Arduino - Home. https://www.arduino.cc/, 2020. [Online; accessed 03-August-2020].

[11] Chujowy CTF. https://chujowyc.tf/, 2020. [Online; accessed 24-July-2020].

[12] ChujowyCTF - Ford CPU — Ethan Wu. https://ethanwu.dev/blog/2020/07/16/chujowy-ctf-ford-cpu/, 2020. [Online; accessed 24-July-2020].

[13] ChujowyCTF 2020 - Ford CPU I & II - Daniel Brodsky. https://www.danbrodsky.me/writeups/chujowyctf2020-fordcpu/, 2020. [Online; accessed 24-July-2020].

[14] CTFTime.org / Google Capture The Flag 2020 / basics / Writeup. https://ctftime.org/writeup/23035, 2020. [Online; accessed 26-August-2020].

[15] CWE-79: Improper Neutralization of Input During Web Page Generation ('Cross-site Scripting'). https://cwe.mitre.org/data/definitions/79.html, 2020. [Online; accessed 21-July-2020].

[16] CWE-89: Neutralization of Special Elements used in an SQL Command ('SQL Injection'). https://cwe.mitre.org/data/definitions/89.html, 2020. [Online; accessed 21-July-2020].

[17] CWE-94: Improper Control of Generation of Code ('Code Injection'). https://cwe.mitre.org/data/definitions/94.html, 2020. [Online; accessed 21-July-2020].

[18] Cyber NYC. https://cyber-nyc.com/, 2020. [Online; accessed 03-August-2020].

[19] Design Automation Conference. https://www.dac.com/, 2020. [Online; accessed 24-July-2020].

[20] Google CTF - Build your future with Google. https://buildyourfuture.withgoogle.com/events/ctf/, 2020. [Online; accessed 24-July-2020].

[21] Hack@DAC2020. https://hackat.events/dac20/, 2020. [Online; accessed 24-July-2020].

[22] Hardware Wallet - State-of-the-art security of crypto assets — Ledger. https://www.ledger.com/, 2020. [Online; accessed 03-August-2020].

[23] hardwear.io — Hardware Security Conference & Training - Netherlands, Germany & USA. https://hardwear.io/, 2020. [Online; accessed 23-July-2020].

[24] Multiple reconfigurable silicon in a single package. https://www.secube.eu, 2020. [Online; accessed 03-August-2020].

[25] Oustanding security diagnostic and support - Riscure. https://www.riscure.com/, 2020. [Online; accessed 23-July-2020].

[26] R. Baldoni, R. De Nicola, and P. Prinetto. *Il Futuro della Cybersecurity in Italia: Ambiti Progettuali Strategici*, chapter 4, pages 80–86. Consorzio Interuniversitario Nazionale per l'Informatica - CINI, 2018. ISBN: 9788894137330.

[27] S. Bratus. What hackers learn that the rest of us don't: Notes on hacker curriculum. *IEEE Security Privacy*, 5(4):72–75, 2007.

[28] M. Bushnell and V. Agrawal. *Essentials of electronic testing for digital, memory and mixed-signal VLSI circuits*, chapter 14. Springer Science & Business Media, 2013.

[29] R. S Cheung, J. P Cohen, H. Z Lo, and F. Elia. Challenge based learning in cybersecurity education. In *Proceedings of the International Conference on Security and Management (SAM)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2011.

[30] S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From game design elements to gamefulness: defining" gamification". In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15, 2011.

[31] C. Domas. Hardware backdoors in x86 cpus, 2018.

[32] C. Eagle. Computer security competitions: Expanding educational outcomes. *IEEE Security Privacy*, 11(4):69–71, 2013.

[3] https://www.cyberchallenge.it/

[33] Q. Ge, Y. Yarom, D. Cock, and G. Heiser. A survey of microarchitectural timing attacks and countermeasures on contemporary hardware. *Journal of Cryptographic Engineering*, 8(1):1–27, 2018.

[34] J. Hamari, J. Koivisto, and H. Sarsa. Does gamification work? – a literature review of empirical studies on gamification. In *2014 47th Hawaii International Conference on System Sciences*, pages 3025–3034, 2014.

[35] K. Huotari and J. Hamari. Defining gamification: a service marketing perspective. In *Proceeding of the 16th international academic MindTrek conference*, pages 17–22, 2012.

[36] P. Kocher, J. Horn, A. Fogh, D. Genkin, D. Gruss, W. Haas, M. Hamburg, M. Lipp, S. Mangard, T. Prescher, M. Schwarz, and Y. Yarom. Spectre attacks: Exploiting speculative execution. In *40th IEEE Symposium on Security and Privacy (S&P'19)*, 2019.

[37] P. Kocher, J. Jaffe, and B. Jun. Differential power analysis. In *Annual International Cryptology Conference*, pages 388–397. Springer, 1999.

[38] P. C. Kocher. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In *Annual International Cryptology Conference*, pages 104–113. Springer, 1996.

[39] K. Leune and S. J Petrilli Jr. Using capture-the-flag to enhance the effectiveness of cybersecurity education. In *Proceedings of the 18th Annual Conference on Information Technology Education*, pages 47–52, 2017.

[40] Y. Li, M. Chen, and J. Wang. Introduction to side-channel attacks and fault attacks. In *2016 Asia-Pacific International Symposium on Electromagnetic Compatibility (APEMC)*, volume 01, pages 573–575, May 2016.

[41] M. Lipp, M. Schwarz, D. Gruss, T. Prescher, W. Haas, A. Fogh, Horn, S. Mangard, P. Kocher, D. Genkin, Y. Yarom, and M. Hamburg. Meltdown: Reading kernel memory from user space. In *27th USENIX Security Symposium (USENIX Security 18)*, 2018.

[42] J. Mirkovic and P. AH Peterson. Class capture-the-flag exercises. In *2014 USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 14)*, 2014.

[43] C. I. Muntean. Raising engagement in e-learning through gamifica-tion. In *Proc. 6th international conference on virtual learning ICVL*, volume 1, pages 323–329, 2011.

[44] D. Patterson, J. Hennessy, and D. Goldberg. *Computer architecture: a quantitative approach*, volume 2. Morgan Kaufmann San Mateo, CA, 1990.

[45] P. Prinetto and G. Roascio. Hardware security, vulnerabilities, and attacks: A comprehensive taxonomy. In *ITASEC*, pages 177–189, 2020.

[46] K. Xiao, D. Forte, Y. Jin, R. Karri, S. Bhunia, and M. Tehranipoor. Hardware trojans: Lessons learned after one decade of research. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 22(1):6, 2016.

Analysis of Software-Implemented Fault Tolerance: Case Study on Smart Lock

Jakub Lojda, Richard Panek, Jakub Podivinsky, Ondrej Cekan, Martin Krcma, Zdenek Kotasek
Faculty of Information Technology, Brno University of Technology, Centre of Excellence IT4Innovations
Bozetechova 2, 612 66 Brno, Czech Republic
Email: {ilojda, ipanek, ipodivinsky, icekan, ikrcma, kotasek}@fit.vutbr.cz

***Abstract*—In our research, we focus on Fault-Tolerant system design and testing. Recently, we also studied Fault Tolerance against random and deliberate faults of electronic smart locks. In our last research, we tested Software-Implemented Fault Tolerance in the controller of a smart electronic lock. We found out that the most sensitive part is the Instruction Memory, but also that our hardening proved to have only negligible effects on the resulting fault tolerance. In this paper, we extend our experiments and provide further analysis of potential pitfalls when hardening using SIFT. We found out that added hardness may improve resilience to faults. But also, the resilience may be instantly worsened by other factors, such as increased bus traffic. In our research we found out, that our hardening did not improve the resiliency to faults most likely due to the increased bus traffic. This means that it is always important to consider the complete system and also the parts of the system that are easily overlooked.**

***Keywords*—Electronic Lock, Stepper Motor, Fault Tolerance Analysis, Fault Injection, FPGA, IMEM, DMEM, LUT.**

I. Introduction

Recently, the so-called Smart Devices [1] gained their popularity. These include the so-called Smart Electronic Lock [2]. It acts as an ordinary door lock, except it can be unlocked by unordinary means, such as by a gesture on a smartphone [3]. It is obvious that a smart lock is a critical device which must satisfy certain reliability standards.

Reliability in electronic devices can be achieved in two different ways: 1) *Fault Avoidance* (FA) [4], which selects from reliable components to build the system. 2) *Fault Tolerance* (FT) [5], on the contrary, changes the structure of the system, so a component failure is not observable on the system behavior. From the FT, the so-called *Software-Implemented Fault Tolerance* (SIFT) [6] is derived, which changes SW code structure to increase its reliability.

Our research focuses on FT design and evaluation. It is important to intensively test FT systems to ensure their quality. For this purpose, the so-called Fault Injection can be used, which intentionally introduces faults into the system. During this, the system is observed and its behavior is evaluated. We hardened and evaluated an electronic lock controller in our previous paper [7]. For the purpose of evaluation, its processor was implemented in *Field Programmable Gate Array* (FPGA), which offered us the possibility to inject faults at run time. Fault injections into the *Instruction Memory* (IMEM), *Data Memory* (DMEM) and the CPU logic itself were evaluated. The results indicated, that the most sensitive is the IMEM. Our tests were held on three different programs, out of which two contained SIFT. The data showed, however, that our SIFT methods did not prove to be beneficial. In opposite, our SIFT made the systems more vulnerable. And this paper focuses on the analysis and explanation of such behavior, as we believe that identifying and avoiding such anomaly is useful in the following research. In this paper, we add a new set of experiments and analyze three additional aspects that are related to the mentioned anomaly. These include: 1) compiler program code optimization; 2) accuracy of DMEM occupied bytes detection; and 3) increase of CPU internal bus transfer rate, which could possibly explain the anomaly.

Security and safety of smart electronic locks are studied in the literature. For example authors of [8] present survey on various identification systems that are usually used in smart locks. Another paper [9] presents a detailed analysis of the security of a specific commercially available smart lock. New SIFT methods can also be found in the literature. Authors of [10] introduce and evaluate a method utilizing unused resources to implement SIFT on the Itanium 2 CPU, which is the *Explicitly Parallel Instruction Computing* (EPIC) processor. Another paper [11] presents an analytic method to evaluate reliability of multi-computer SIFT systems.

This paper is organized as follows. Electronic locks structure with discussion about their reliability is presented in Section II. Evaluation platform for monitoring faults impacts in electro-mechanical systems is presented in Section III. Experimental evaluation of faults injected into SW controller program (stored in IMEM) and run-time data (stored in DMEM), alongside with injection into HW logic in LUTs, is presented in Section IV. Section V presents the analysis of our results and concludes the paper.

II. Electronic Locks

Smart electronic lock is a relatively complex device that uses the latest technologies of nowadays. It typically consists of three parts (modules) [12]: 1) Control Module; 2) Motor Module; and 3) I/O Module. The management of the entire lock is provided by the Control Module which performs a number of computational extensive operations, therefore, it is typically realized by a processor. The mechanical part of the lock consists of the Motor Module, which can be realized by

978-1-7281-9899-6/20/$31.00

various drives that manipulate the lock. In our research, we focus on a stepper motor, which is very often used in smart locks as the motor [13]. The stepper motor has its rotation divided into several equal steps, allowing precise control of the position of the rotation by means of input pulses. The I/O Module is used for a communication and performs mainly the communication with interfaces such as Wi-Fi or Bluetooth.

In our research, we focus on the change of the processor data. This can be a program change – another instruction sequence is executed, or a data change – other values are used. The injection of faults into the processor may result in unexpected and unwanted behavior of the smart lock and consequently property damage. The fault can be induced naturally from environment via charged particle or through attacker which intends to change the data in memory by electromagnetic interference or by specific material that secretes these particles. The fault can also occur when attacker mechanically damages the smart lock, its circuit board or another component.

When data are corrupted, the lock can be unlocked if incorrect authentication is performed or can stay in the lock state when unlocking with the correct credentials. It may also happen that the lock is not really locked when the lock is requested. Anyone will have access to a permanently unlocked door. However, there may also be a failure that occurs only in a certain situation, i.e. only in a certain state of the lock. Such a fault is very difficult to be detected and it is not entirely clear when and what behavior will occur in the fault. Therefore, in this research, we focus on the impact of these faults in the smart lock on the processor.

III. Evaluation Platform

In our previous work we introduced a platform for fault tolerance evaluation [14]. This platform is based on functional verification principles combined with faults injection into an FPGA. Functional verification is based on the simulation of a verified system and monitoring its outputs and comparing it to a reference data after feeding predefined inputs to the system. We used this principle for our purposes, however we implement the verified system into an FPGA which allows us to easily inject faults to the system and evaluate their effects.

The platform capabilities were demonstrated on an example of a robot searching for its way through a maze. It was an FPGA controlled simulated system aimed to experimentally evaluate the platform, however the platform was designed to be scalable and able to evaluate any system controlled by an FPGA. It offers a convenient way to evaluate faults effect on the controller and the stepper motor of an electronic lock. The experimental results of this work are based on our platform.

To successfully use the platform, we have been forced to modify application specific components of the platform. It is necessary for the controller to be implemented in an FPGA. It is vital to establish a proper communication line between the control unit operating in the FPGA and the software simulating the stepper motor running on the different computing platform. For these purposes we use MATLAB and the Simulink [15] software, specifically the Simscape [16] library. In this case, the communication is realized via Ethernet. During the evaluation, the platform monitors the faults effect not only on the controller but also on the mechanical part of the system - the stepper motor. To be able to do this, the platform utilizes a simulation with autonomous analysis of the motor behavior. It is also vital to choose a proper injection strategy as it has a significant impact on results quality.

IV. Experiments and Results

To test faults in the CPU and its memories during their operation, we use implementation of the MSP430 CPU for the FPGA, called the NEO430 [17]. The three original programs from our previous research [7] were extended with one new program, to isolate the effects of our SIFT methods. Further, we changed the naming of our previous programs, as these might be confusing in the context of our new analysis. Actual SIFT modifications made to the original code are shown in the Activity Diagram in Figure 1.

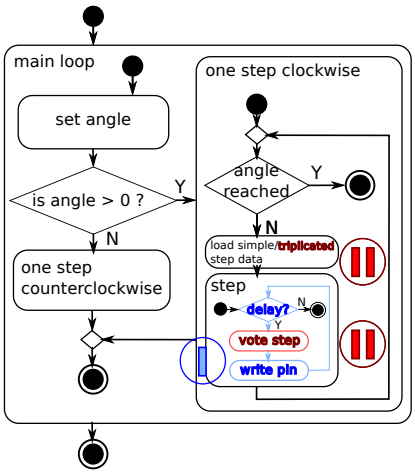


Figure 1: UML Activity diagram of the program with two modifications: ”I”) writes signals to output pins during the delay; ”II”) triples stored data and adds voting.

The program variants include: **1) Original Variant** (i.e. Variant O) – the unmodified control program; **2) Variant I** – it propagates signals to the output pins during the delay function execution; **3) newly added Variant II** – the motor excitation data are stored in three copies in the IMEM and are voted before their usage; **4) Variant I+II** – combines both the modifications I and II. For the implementation, we use the Xilinx Virtex 5 FPGA and synthesize the logic using the *Integrated Synthesis Environment* (ISE) 14.7. Again, we examined two injection strategies: 1) the single; and 2) the multiple fault injection.

A. Single Fault Injection Experiments

During the single fault experiment, one bit flip fault is injected before the CPU clock signal is enabled. For single experiments, injections into utilized bytes of IMEM and the CPU were examined independently. CPU injection is approximated through a bit flip in the occupied *Look-up Tables* (LUTs) of the CPU FPGA implementation. The faults were selected uniformly-at-random and 6,000 runs were performed for the CPU, while 2,000 runs for the IMEM. The results are shown in Table I. The first part of Table I classifies failures into ***Stuck*** – the motor stopped too early; ***Timeout*** – the motor did not

stop during the predefined interval of 220 s; and ***Mismatch*** – wrong data were observed on the output pins. The right part of Table I classifies the cases that achieved the correct angle although the electronic showed errors on its outputs. Although injection into the IMEM shows only slightly better results for the hardened programs I, II and I+II, the CPU injection shows the opposite trend. As can be observed, the hardening was apparently worsened by an unpredicted phenomenon.

TABLE I: The results of single injection experiments with failures classification; ”O”, ”I” and ”I+II” published in [7], extended with ”II”.

	Electronic Failure				*Mechanic OK (Out of Electronic Failed Runs)*			
	Total [%]	*Stuck* [%]	*Timeout* [%]	*Mis-match* [%]	*Total* [%]	*Stuck* [%]	*Timeout* [%]	*Mis-match* [%]
CPU ”O”	5.3	4.5	0.4	0.4	7.3	3.2	0.0	4.4
IMEM ”O”	36.7	15.6	14.0	7.0	20.3	1.4	1.0	18.0
CPU ”I”	5.9	5.1	0.5	0.3	8.4	3.9	0.7	3.9
IMEM ”I”	35.2	15.5	14.4	5.4	21.2	4.8	3.7	12.6
CPU ”II”	6.1	5.6	0.4	0.2	8.4	6.5	0.3	1.6
IMEM ”II”	34.6	16.0	11.9	6.7	24.2	3.0	2.6	18.5
CPU ”I+II”	6.2	5.5	0.5	0.2	10.4	5.1	0.3	2.4
IMEM ”I+II”	34.0	17.3	9.6	7.1	25.0	3.8	1.6	19.6

It is important to evaluate the mechanics behavior, too. In Figure 2, the final number of motor rotations for experiment runs in which the electronic failed is shown in a box plot chart. The desired 12.4 rotations is highlighted by the blue line.

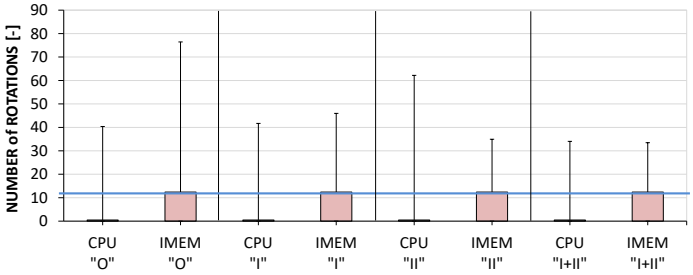


Figure 2: Box plot chart with final rotation for single injection; ”O”, ”I” and ”I+II” published in [7], extended with ”II”.

As can be observed, in most cases, the CPU injection caused that the motor did not start to rotate at all. For the IMEM injection, the hardened programs have the maximum angle slightly closer to the required value of 12.4 rotations.

B. Multiple Fault Injection Experiments

We made equivalent experiment with the multiple injection strategy. At the beginning of each run, the processor was started. After first 10 s, one bit flip was injected every 5 s. A run was active until the motor stopped on the required angle (usual duration of 80 s) or a timeout of 220 s was achieved. The faults were selected uniformly-at-random and 6,000 runs were held for the CPU target; 2,000 runs for the IMEM and 500 runs for the DMEM. These experiments also include injection into the DMEM, which is not meaningful for single injection strategy, as the DMEM contents is built during run time. The occupied DMEM was detected based on circa 1,000 DMEM read backs, before the injection experiments were started. Through the analysis of the read backs, we obtained the memory utilization map. Data are presented in Table II. The meaning of the columns is equivalent to the single experiments.

TABLE II: The results of multiple injection experiments with failures classification; ”O”, ”I” and ”I+II” published in [7], extended with ”II”

	Electronic Failure				*Mechanic OK (Out of Electronic Failed Runs)*			
	Total [%]	*Stuck* [%]	*Time-out* [%]	*Mis-match* [%]	*Total* [%]	*Stuck* [%]	*Time-out* [%]	*Mis-match* [%]
CPU ”O”	71.3	14.4	24.9	32.1	16.0	2.0	1.5	12.5
IMEM ”O”	99.1	41.1	27.1	30.9	19.7	1.8	0.2	17.7
DMEM ”O”	91.8	9.0	15.4	67.4	13.1	0.0	0.0	13.1
CPU ”I”	70.1	13.6	30.1	27.0	16.3	2.2	3.2	10.9
IMEM ”I”	98.4	31.8	31.4	35.3	31.7	4.4	0.2	27.1
DMEM ”I”	92.8	11.6	15.6	65.6	13.8	0.0	0.0	13.8
CPU ”II”	68.7	16.0	40.3	12.4	14.8	3.4	3.9	7.6
IMEM ”II”	99.0	53.9	16.4	28.8	43.4	3.4	0.0	18.3
DMEM ”II”	99.2	60.6	2.6	36.0	6.9	0.0	0.0	6.9
CPU ”I+II”	89.0	19.6	23.4	46.0	12.1	1.9	1.0	9.2
IMEM ”I+II”	99.7	40.1	22.5	37.1	27.3	3.9	0.2	23.3
DMEM ”I+II”	95.4	34.2	3.0	58.2	17.2	0.0	0.0	17.2

As can be seen, generally the highest sensitivity has the IMEM. This is because it is often read, a change in its content alters the program behavior and the IMEM is never written to. This is why a fault in the IMEM has no possibility to eventually rewrite (i.e. repair) during the program run time. Also, as can be observed, we believe the higher DMEM occupancy worsens the results of the experiments with DMEM. We also believe that the higher utilization of the internal CPU bus for the ”I+II” program causes the deviation of the CPU ”I+II”, which has significantly worse results.

We also monitored mechanic part for the multiple injections. The final number of rotations for runs in which the electronic failed can be seen in box plot chart in Figure 3.

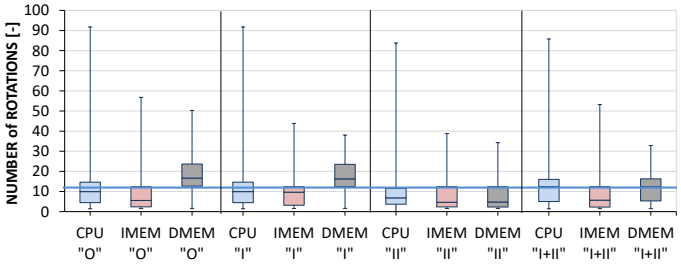


Figure 3: Box plot chart with final rotation for multiple injection; ”O”, ”I”, ”I+II” published in [7], extended with ”II”.

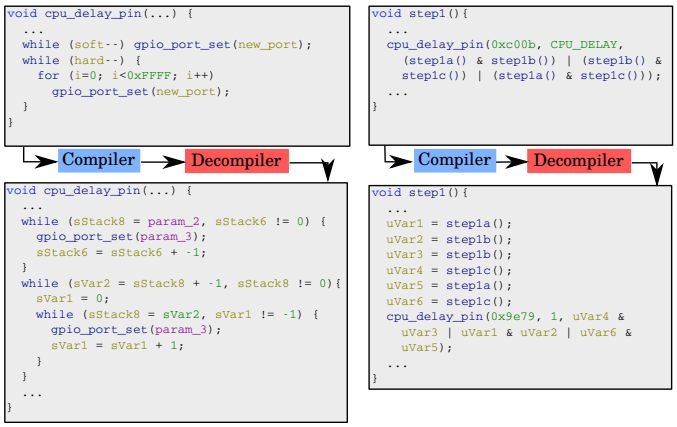
As can be observed, the rotation for ”I” compared to ”O” is very similar for the CPU and DMEM targets; for the IMEM, the ”I” is better. The ”II” is worse for all injection targets. However, the ”I+II” has the median closer to the expected rotation for the CPU and DMEM (i.e. better than the ”O”); for the IMEM, the rotation is very similar to the ”O”.

V. Analysis and Conclusions

Although the experimental results are interesting for assessment of faults impact on particular injection targets, the hardening itself did not bring significant improvements. From our point of view, it is very interesting to analyze this anomaly and publish the design error which caused this anomaly. In the following text, three hypothetical reasons are examined.

A. Compile-time Code Optimization

At first, we ensured our SIFT modifications remained in place after the code was compiled, although the lowest possible optimization level was selected. By using the Ghidra tool [18], we decompiled our binary programs for the NEO430. Code snippets of modifications "I" and "II" can be seen in Figure 4. As can be seen, the hardening remains in the binary program.



Code Modification I (Data are Continually Propagated during the Dealy Operation)

Code Modification II (Data are Stored Three Times in the IMEM)

Figure 4: Original vs. decompiled program code snippets for both of the modifications.

B. Accuracy of our DMEM Occupancy Detection

For our research, we developed the detector of DMEM address occupancy. This significantly accelerates the evaluation, as the average DMEM occupancy for our programs is 1.9 %. However, if a dynamic memory allocation is in place, the occupied addresses may be fragmented all over the address space. And for such cases, our detection method is not suitable. To evaluate the suitability of this method, we created heat maps of memory bytes occupancy. It is obvious that a high *temperature* on a small number of cells indicates a better suitability. On the contrary, a low *temperature* on a high number of memory addresses indicates the occupied cells are scattered. The heat maps can be seen in Figure 5.

It is obvious that the method is suitable for our programs, as the heat maps indicate occupancy of a few cells with a high probability of them being occupied.

C. CPU Internal Bus Traffic

The NEO430 is a 16-bit processor. It uses the internal host bus to communicate with its numerous components. Considerable amount of space is occupied by bus controllers

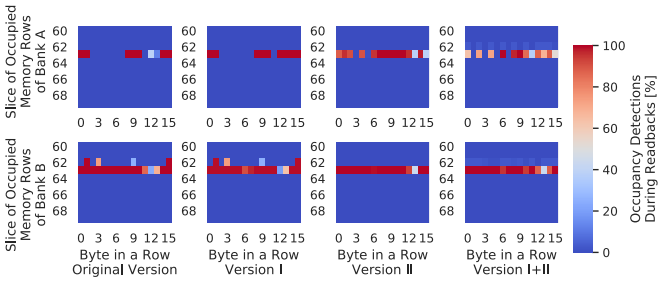


Figure 5: Heat maps of DMEM occupancy for each of the four programs.

of the components. It is, therefore, hypothetically possible that the added SIFT increased the transmission rate on the internal bus. This potentially enlarges the space for fault manifestation. For each program, we measured the number of read and write transactions. As the NEO430 distinguishes between both the bytes in a 16-bit write transaction, these were monitored independently (i.e. as the Byte 0 and Byte 1). We also monitored the amount of data transferred. The results are shown in Table III.

TABLE III: Bus traffic for each of the four program versions

	Program	***Original***	***Version I***	***Version II***	***Version I+II***
Transactions	*Read*	1556.6×10^6	1892.4×10^6	1529.5×10^6	1894.7×10^6
	Write Byte 0	29 143	287.1×10^6	29 768	287.1×10^6
	Write Byte 1	29 143	287.1×10^6	31 876	287.1×10^6
Data	*Read*	2968.9 MiB	3609.5 MiB	2917.3 MiB	3613.9 MiB
	Written	0.055 MiB	547.6 MiB	0.058 MiB	547.6 MiB

As can be observed, the high number of read transactions is caused by reading program instructions from the IMEM. Furthermore, the repeated propagation of results to the output pins (i.e. the modification "I") significantly increased the write transactions number for corresponding programs. These results indicate that, at least for the CPU experiments, the added hardness was partially cancelled by making the program more vulnerable due to increased bus traffic.

To conclude this paper, we found out that added hardness may improve resilience to faults. But also, the resilience may be instantly worsened by other factors, such as increased bus traffic. In our research we found out, that our hardening did not improve the resiliency to faults due to the increased bus traffic. This means that it is always important to consider also the parts of the system that are easily overlooked. And it is necessary to search for other critical points for the FT.

Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project IT4Innovations excellence in science – LQ1602, the Brno University of Technology under number FIT-S-20-6309 and the JU ECSEL Project SECREDAS (Product Security for Cross Domain Reliable Dependable Automated Systems), Grant agreement No. 783119.

REFERENCES

[1] C. Salzmann, S. Govaerts, W. Halimi, and D. Gillet, "The smart device specification for remote labs," in *Proceedings of 2015 12th International Conference on Remote Engineering and Virtual Instrumentation (REV)*. IEEE, 2015, pp. 199–208.

[2] Y. T. Park, P. Sthapit, and J.-Y. Pyun, "Smart digital door lock for the home automation," in *TENCON 2009-2009 IEEE Region 10 Conference*. IEEE, 2009, pp. 1–6.

[3] C. Lee, Y. Chung, T. Shen, and K. Weng, "Development of electronic locks using gesture password of smartphone base on rsa algorithm," in *2017 International Conference on Applied System Innovation (ICASI)*, 2017, pp. 449–452.

[4] J.-C. Geffroy and G. Motet, *Design of Dependable Computing Systems*. Kluwer Academic Publishers, 2002.

[5] I. Koren and C. M. Krishna, *Fault-Tolerant Systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007.

[6] L. Pullum, *Software Fault Tolerance Techniques and Implementation*, ser. Artech House computing library. Artech House, 2001. [Online]. Available: https://books.google.cz/books?id=hqXvxsO5xz8C

[7] J. Lojda, R. Panek, J. Podivinsky, O. Cekan, M. Krcma, and Z. Kotasek, "Hardening of Smart Electronic Lock Software against Random and Deliberate Faults," in *Paper accepted for presentation at Digital System Design (DSD), 2020, 23th Euromicro Conference*. IEEE.

[8] R. S. Divya and M. Mathew, "Survey on various door lock access control mechanisms," in *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, 2017, pp. 1–3.

[9] E. Knight, S. Lord, and B. Arief, "Lock picking in the era of internet of things," in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, 2019, pp. 835–842.

[10] G. A. Reis, J. Chang, N. Vachharajani, R. Rangan, and D. I. August, "Swift: software implemented fault tolerance," in *International Symposium on Code Generation and Optimization*, March 2005, pp. 243–254.

[11] D. R. Avresky, S. J. Geoghegan, and Y. Varoglu, "Evaluation of software-implemented fault-tolerance (sift) approach in gracefully degradable multi-computer systems," *IEEE Transactions on Reliability*, vol. 55, no. 3, pp. 451–457, 2006.

[12] Y. T. Park, P. Sthapit, and J. Pyun, "Smart digital door lock for the home automation," in *TENCON 2009 - 2009 IEEE Region 10 Conference*, Jan 2009, pp. 1–6.

[13] G. K. Verma and P. Tripathi, "A digital security system with door lock system using RFID technology," *International Journal of Computer Applications*, vol. 5, no. 11, pp. 6–8, 2010.

[14] J. Podivinsky, O. Cekan, J. Lojda, M. Zachariasova, M. Krcma, and Z. Kotasek, "Functional Verification based Platform for Evaluating Fault Tolerance Properties," *Microprocessors and Microsystems*, vol. 52, pp. 145 – 159, 2017.

[15] MathWork®, "MATLAB and Simulink," https://www.mathworks.com/, 2018, accessed: 2019-03-20.

[16] MathWork®, "Stepper motor," https://www.mathworks.com/help/physmod/sps/powersys/ref/steppermotor.html, 2019, accessed: 2019-03-20.

[17] S. Nolting, "NEO430 Processor," https://github.com/stnolting/neo430, 2018.

[18] National Security Agency, "Ghidra - Software Reverse Engineering Framework," https://www.nsa.gov/resources/everyone/ghidra/, 2020, accessed: 2020-06-20.

An Indoor Smart Lamp For Environments Illuminated Day Time

Ayşe Nur Cihan
Department of Computer Engineering Faculty of Technology
Selcuk University
Konya, Turkey
cihanaysenurr@gmail.com

Gül Nihal Güğül, IEEE Member
Department of Computer Engineering, Faculty of Technology
Selcuk University
Konya, Turkey
gul.gugul@selcuk.edu.tr

***Abstract*— Lighting energy consumption constitutes % 9-25 of final consumption in buildings. In addition, lights are kept open during daytime in some buildings such as public institutions and schools.**

The main purpose of this study is to design a remotely controlled smart lamp that measures the illumination of the environment and adjusts illumination level of lamp according to demand in order to decrease the electricity consumption due to lighting during daytime. Also, in this study energy savings that can be achieved with the usage of this lamp in a school building in analyzed. Many studies are conducted to develop smart lamps, however to the authors best knowledge a remotely controlled autonomous smart lamp by Wi-Fi module and a software developed in C# interface, to update required minimum illumination level of the room is not found in literature. By using the developed software minimum illumination levels of the rooms of a whole building can be managed remotely which is useful in cases such as changing the purpose or color of the rooms or changing minimum level seasonally.

System consists of Arduino, BH1750 light sensor, strap LEDs, transistors and Wi-Fi module. Daylight illumination level is measured with BH1750. Smart lamp changes its illumination level by taking the difference of measured illumination data from minimum illumination level data, if the measured illumination level is lower than minimum level. In order to analyze the energy saving that could be achieved, illumination level of the indoor and outdoor environment of Faculty of Technology of Selcuk University is measured at 30 minutes' intervals for one day. Energy saving to be obtained in case of using developed smart lamp is calculated as 1747 kWh/year only for daytimes.

Keywords—smart lamp, lighting software, remote control, energy saving

I. INTRODUCTION

Energy consumption for lighting in US is 25% in commercial buildings, 12% in residential buildings [1] according to US DOE 2009 data, whereas 9% in residential buildings in Turkey according to a study conducted in 265 residential buildings in Ankara, Turkey [2]. U.S. Energy Information Administration Residential Energy Consumption Survey shows that lighting in residential buildings reduced to 10% by 2015 in US [3] mainly due to LED lamps. However, energy consumed for lighting in buildings is still high enough to investigate reduction methods and develop more efficient lamps.

The correct amount of illumination in indoor spaces increases vision, protects eye health, affects working performance and is important in terms of mental health [4]. For this reason, correct amount of illumination is very significant especially in school buildings. Many calculation techniques are used to calculate the minimum required illumination for indoor environments. Design and simulation softwares for lighting are available to facilitate the use of these techniques. The most common commercial lighting design tools are DIALux and Relux. DIALux is one of the world's leading programs developed to plan, calculate and visualize the amount of light and is available free [5]. Relux is another free lighting account tool developed in Germany. Calculation is made by considering the number of luminaires to be used, indoor illumination level value, the color of the environment and three-dimensional objects [6]. Many studies are conducted by using these commercial software ([7], [8]).

Besides the commercial software, there are also studies conducted to develop software and determine the required lighting power in the literature. In the PhD thesis completed at Istanbul Technical University Energy Institute, a calculation method has been developed in order to determine the lighting energy saving potential and a tool (bep/ETA) has been developed by using MS Excel to use the developed method [9]. In a study conducted at Usak University, a software that performs indoor lighting analysis was designed by using the C # programming language. This software has been developed as an educational software to be used in schools where lighting education is provided [10]. At present, a wide range of day lighting simulation software's are available in the industry such as ECOTECT, Energy Plus and Radiance [11].

A dynamic lighting design software is developed in scope of E.U. funded research project, concerning ways of designing dynamic luminous environments in outdoor environments. D.L.D. software is capable of controlling lighting installations of urban public spaces [12]. Also softwares are developed not just for buildings but also for smart cities. Intelligent Street Lighting Software is developed for lighting control in smart cities in Spain to control public lighting [13]. There are also some tools for smart cities such as Lites (has temperature sensors, ambient light, power, motion detection), CityLight (has remote management of lighting, fault detection and planning lighting patterns manually) and Tvilight (regulates the lighting based on presence sensors and maintains minimum illumination in inactive hours) [13]. Design of a smart indoor lighting is developed in NBN Sinhgad School of Engineering by using motion and light sensor based on fuzzy logic controller. In auto-mode, lights are switched on with motion and controls light intensity depending on available natural light [14]. Another study introduced design method of PWM control and detection system for the power intensity of white LED based on a microcomputer [15]. A smart LED lighting system is developed that combines the LED lighting technology with infrared sensing technology, photoelectric detection technology and intelligent control technology. The

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

system controls intelligently the illumination of the lamp for daily lighting needs and played an important role in saving energy [16]. A smart lamp is developed with a motion sensor in a study conducted in North-Eastern Federal University based on the Arduino and controls using a smartphone [17].

There are studies conducted to develop smart lamps, however to the authors best knowledge a smart lamp with a software developed in C# interface in order to update required minimum illumination intensity is not found in literature. By using the developed software, the necessary minimum illumination levels of the rooms of a whole building can be managed remotely and results is significant energy saving during daytime in cases such as changing the purpose or color of the rooms or changing it seasonally.

In the proposed study, a software is developed that calculates the required minimum illumination intensity in a room by taking into account the size of the environment, reflection coefficients, lamp type, power, pollution factor and also indoor colors. In addition to developed software, a lamp is designed that receives data from software by Wi-Fi module. Then lamp measures the illumination in the relevant room. By subtracting the measured data from required minimum data, lamp determines the instant required illumination level and adjusts power of the lamp according to instant required illumination during daytime if the illumination level inside the room is lower than minimum necessary illumination level.

In addition to developed software and lamp, illumination level of the outdoor and interior environment are measured in classrooms facing the east and west directions for one day for 30 minutes' interval in Faculty of Technology of Selcuk University to calculate the rate of indoor illumination to outdoor illumination. Finally, the energy saving to be obtained in case using the developed lamp in this building is calculated for one year by using annual hourly illumination data.

II. METHODOLOGY

In this study a software is developed in order to calculate the required minimum illumination level in the room. Then calculated value is sent to smart lamp by Wi-Fi module. Smart lamp measures the illumination of the room by using BH1750 sensor and subtracts the measured value from the value sent by software. Finally, smart lamp turns on the required number of LEDs in order to fulfill he illumination demand in the room.

In this section firstly developed software is described. Then smart lamp is explained in detail. Finally, the calculation method of the energy saving in case of using developed lamp in a school building is given.

A. Development of Lighting Software

Light intensity is the measure of the light flux emitted from the light source. Level of illumination is the sum of the luminous flux per unit surface and indicates the level of light that the light source gives in every direction. Unit of illumination level is lux [4].

The software in this study is developed in order to calculate the minimum illumination level required in a room. At the development stage of the software in C#, the values in Table 1 and Table 2, and the equations (1) to (4) are used. Tables were created using the SQL database. Flow chart of lighting software is given in Fig. 1.

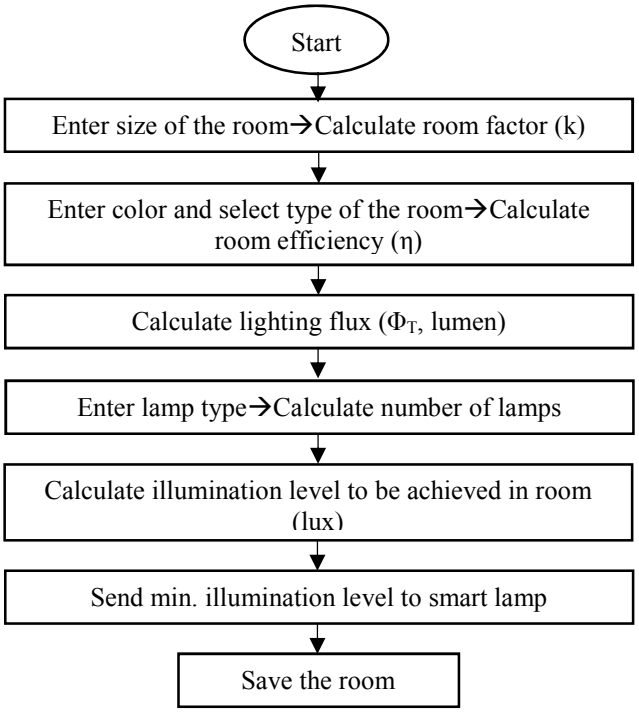


Fig. 1. Flow chart of lighting software

Firstly, room factor (k) is calculated by (1), to calculate the required minimum illumination in a dark environment, [18].

$$k = \frac{a \times b}{H \times (a+b)} \quad (1)$$

In this equation;
k : Room factor
a : Short edge length of the room, meters
b : Long edge length of the room, meters
H : Height between lamp and work surface, meters

Table for reflection coefficients of colors are generated in SQL database [18]. Efficiency value (η) depends on physical properties of the room such as color and room factor [18]. Room efficiency (η) is determined by matching the row of related room factor (k) with the column of related reflection coefficient in Table 1.

TABLE I. ROOM EFFICIENCIES BY ROOM FACTOR [18]

	Reflection coefficients									
Ceil	0,8				0,5				0,3	
Wall	0,5		0,3		0,5		0,3		0,1	0,3
Floor	0,3	0,1	0,3	0,1	0,3	0,1	0,3	0,1	0,3	0,1
k	**Room efficiency, ŋ**									
0,6	0,2	0,2	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,1
0,8	0,3	0,2	0,2	0,2	0,2	0,2	0,2	0,1	0,1	0,1
1,00	0,3	0,3	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,2
1,25	0,4	0,3	0,3	0,3	0,3	0,3	0,2	0,2	0,2	0,2
1,50	0,4	0,4	0,3	0,3	0,3	0,4	0,3	0,3	0,2	0,2
2,00	0,5	0,4	0,4	0,4	0,4	0,3	0,3	0,3	0,3	0,3
2,50	0,5	0,4	0,5	0,4	0,4	0,4	0,4	0,3	0,3	0,3
3,00	0,5	0,5	0,5	0,4	0,4	0,4	0,4	0,4	0,3	0,3
4,00	0,6	0,5	0,5	0,5	0,5	0,4	0,4	0,4	0,4	0,3
5,00	0,6	0,5	0,6	0,5	0,5	0,4	0,5	0,4	0,4	0,4

Minimum illumination level required in the rooms depends on the purpose of the room and 300 for classrooms [18]. In order to calculate the luminous flux (2) is used [18].

$$\Phi_T = \frac{d \times E \times a \times b}{\eta} \quad (2)$$

In this equation;
Φ_T : Luminous flux required for lighting, lumen
d : Pollution Factor
E : Minimum illumination level of selected room type
η : Room lighting efficiency

Number of lamps to be used is calculated by (3) [18].

$$Z = \frac{\Phi_T}{\phi} \quad (3)$$

In this equation;
Z : Total number of lamps
ϕ : Luminous flux value of one lamp to be used, lumen

After determining the lamp type and number of lamps, the illumination level to be achieved is calculated with (4) [18].

$$E_s = \frac{\phi \times Z \times \eta}{d \times a \times b} \quad (4)$$

In this equation;
E_s : Illumination level, lux
d : Pollution factor

Illumination level to be achieved in the room is sent to smart lamp by wi-fi. This value is a constant value. The main purpose of this software is to change this value whenever the user wants. User can change the value according to seasons or purpose of the room. Also pollution factor of the lamps can be updated once a year by the user.

B. *Development of Smart Lamp*

In this section developed smart lamp is described in detail. System structure of smart lamp is given in Fig. 2. Lamp is controlled by Arduino. Wi-Fi module is used in Arduino to get data from software. Lamp is composed of a strap led.

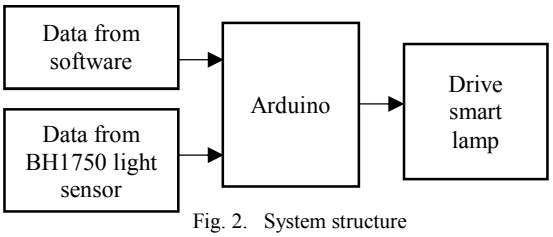


Fig. 2. System structure

BH1750 light intensity sensor is used to measure illumination value in lux. Minimum required light intensity data is sent from software to the lamp. Instead of using strap LEDs in this lamp for meters, the strap led is divided into parts because it is a prototype and each part is considered as 1 meter. If 2 pieces of LEDs are on, it means 2 meters of LEDs are on. The lighting product obtained by arranging 72 new generation light 5630 chip LEDs on 1-meter. Calculated data in software is sent to the Arduino via Wi-Fi with the "send data" button.

Eight transistors are used for switching LEDs. A transistor is connected for each strap led part. LEDs are powered by 12V and Arduino is powered by 5V. It has switched to control 12V led with 5V light signal. Circuit diagram of smart lamp is given in Fig. 3.

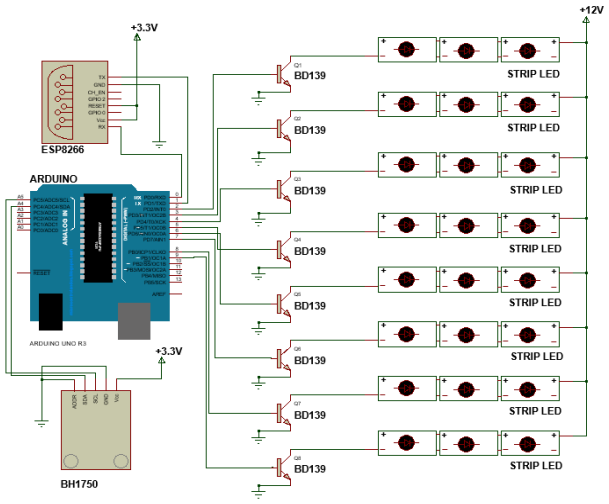


Fig. 3. Circuit diagram of the developed smart lamp

C. *Calculation method of the energy to be saved*

Faculty of Technology of Selcuk University is shown in Fig. 4 and Fig. 5.

Fig. 4. Faculty of Technology, Selcuk University

Fig. 5. Classroom area in each floor

Electricity that could be saved in case of using the developed smart lamp in classrooms of Faculty of Technology of Selcuk University is calculated. There are classrooms in 4 floors as seen in Fig. 4. and classrooms in each floor is nearly 550 m^2, of which's 275 m^2 is faced to east and 275 is faced west as seen in Fig. 5. Firstly, illumination level in the desk level in middle of classrooms faced to west and east is measured and the illumination level in outdoor is measured in 30 minutes' intervals from 9:00 am to 17:00 pm during one day. This data gives us the illumination percentage that penetrates into the classroom and assumed to be constant during one year. Hourly global illumination level for one year of Konya, Turkey is obtained from "Climate.OneBuilding" [19]. By using the illumination percentage, the illumination level inside the classrooms is calculated annually during

daytime. This value gives the value that could be measured by BH1750 sensor. Finally subtracting the BH1750 value from minimum illumination level, the power of the smart lamp is calculated by using (5) hourly.

$$P = E \times A_c / \eta_e \quad (5)$$

In this equation;
P : Power of lamp, Watt
E : Calculated illumination
A_c : Area of all rooms faced to east or west, m2
η_e : Luminous efficacy, Lumen/watt
Luminous efficacy of LED lamps is assumed to be 90 lm/W [20].

III. RESULTS

In this section screen shots of developed software are given. Then developed lamp is shown with features. Finally, the energy to be saved in case of using developed lamp in a school building is calculated.

A. Lighting Software

Minimum required luminous flux of a room is calculated and result is sent to the smart lamp by using developed software. In order to calculate required light intensity firstly a name is given to the room. Then user enters width, length, height and armature height as given in Fig. 6. In next stage color of wall, basement and ceiling that effects efficiency as a reflection factor are provided by user as seen in Fig. 7. Upon the selection of the colors to be used, the lighting efficiency is determined by software. Then room and lamp type and wattage of lamp planned to be used are provided. Finally required minimum illumination is calculated by software and calculated data is sent to smart lamp as seen in Fig. 8.

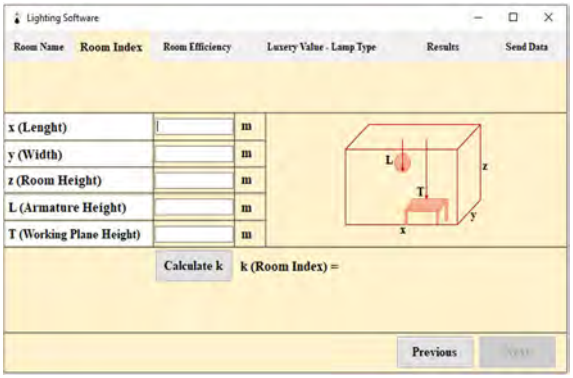


Fig. 6. Calculation of room factor k

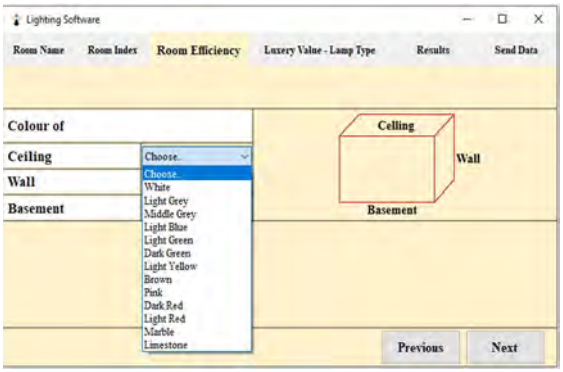


Fig. 7. Colos of interior

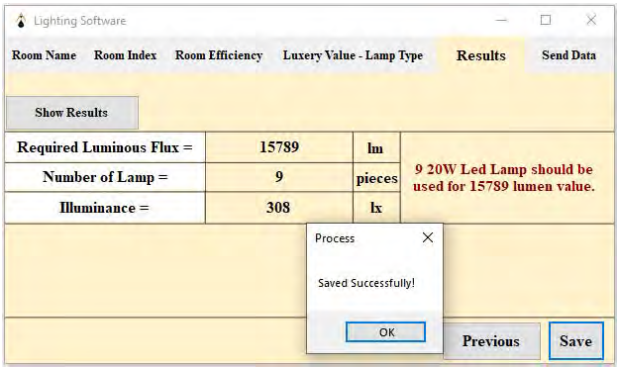


Fig. 8. Results form

Data is sent to smart lamp with Wi-Fi feature of computer. Also project is saved in computer and user can update room properties whenever necessary. Purpose of room may change or color of room can be updated.

B. Smart lamp

Arduino is used in order to control lamp and Wi-Fi module is used to send the data to the C # interface. BH1750 light intensity sensor is used to measure the lux value of the environment. Developed smart lamp is a prototype, so strap led is divided into parts and each part is considered as 1 meter. Eight transistors were used for switching. A transistor is connected for each strap led part. LEDs are powered by 12V and Arduino is powered by 5V. It has switched to control 12V led with 5V light signal. Smart lamp is shown in variable lighting level cases in Fig. 9.

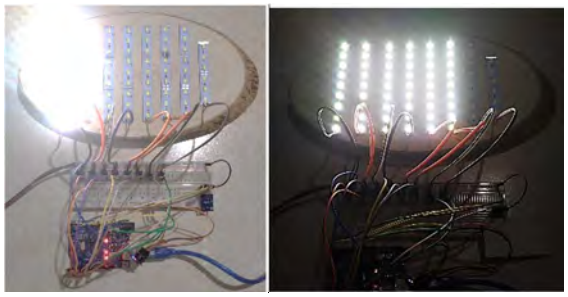


Fig. 9. Smart lamp in various indoor illumination levels

C. Energy to be saved in case of using developed lamp

Amount of energy that could be saved in case of using smart lamp in classrooms of Faculty of Technology of Selcuk University is calculated. There are classrooms in 4 floors of Faculty with floor area nearly 550 m^2, of which's 275 m^2 (23m×12m) faced to east and rest is faced west in each floor.

Minimum illumination level for classrooms is 300 lux [18]. Illumination level of Konya, Turkey is obtained hourly for one year for outdoor environment [19]. Then, illumination level in middle of classrooms at desk level faced to west and east is measured and the illumination level in outdoor is measured in 30 minutes' intervals from 9:00 am to 17:00 pm during one day. Illumination share in indoor for east and west side is calculated by measurement data and are given in Table 2. This data is also obtained by software and sent to smart lamp.

TABLE II. ILLUMINATION LEVEL MEASURED IN 30 MINUTES' INTERVALS FROM 9:00 AM TO 17:00 PM

Time	Out door	Illumination level in class faced to East	East light in indoor, %	Illumination level in class faced to West	West light in indoor, %
09:30	20137	1402	7	181	0,9
10:00	22096	1904	8,6	164	0,7
10:30	25659	2967	11,6	126	0,5
11:00	33684	2110	6,3	157	0,5
11:30	32301	2296	7,1	77	0,2
12:00	14404	772	5,4	75	0,5
12:30	42820	2021	4,7	72	0,2
13:00	44395	1154	2,6	185	0,4
13:30	31553	853	2,7	116	0,4
14:00	27659	534	1,9	129	0,5
14:30	3735	787	21,1	103	2,8
15:00	2066	510	24,7	182	8,8
15:30	2156	515	23,9	170	7,9
16:00	1357	490	36,1	330	24,3
16:30	1722	330	19,2	281	16,3
17:00	2455	272	11,1	190	7,7

By using the illumination share values given in Table 2 and annual weather data, the illumination level inside the classrooms is calculated for one year during daytime separately for east and west sides. This value gives the value that could be measured by BH1750 sensor. Finally subtracting the BH1750 value from minimum illumination level, the lux value of the smart lamp is calculated by using equation (5) in cases of illumination level inside the classroom is lower than 300 lux. If the illumination level is higher than 300 lux, smart lamps are closed, although sometimes they are open in real life. Electricity consumption of all classrooms with smart LED lamp and standard LED lamp is given in Fig. 10.

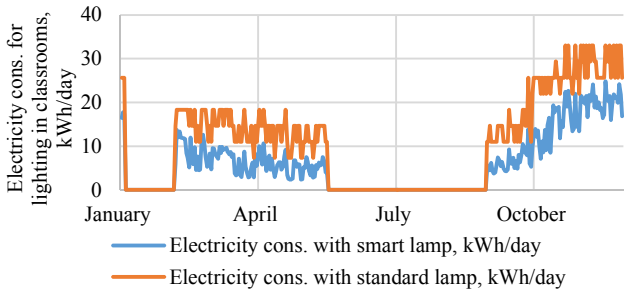


Fig. 10. Electricity consumption of all classrooms with smart and standard LED lamp

As it is seen from Figure 13 electricity consumption for lighting is calculated for non-holiday days. Results showed that 1747 kWh/year electricity could be saved in case of using smart lamp in the classrooms of Faculty of Technology of Selcuk University.

IV. CONCLUSION

In this study a remotely controlled smart lamp is developed in order to decrease lighting electricity consumption during daytime. Smart lamp measures the illumination level of the environment with sensor and adjusts illumination level of lamp. System consists of Arduino, BH1750 light sensor, strap LEDs, transistors and Wi-Fi module. Daylight illumination level in indoor is measured with BH1750. Smart lamp adjusts illumination level by taking the difference of measured illumination data from minimum illumination level data, if the measured illumination level is lower than minimum level. Energy saving that would be achieved by using developed lamp is analyzed in Faculty of Technology of Selcuk University. Annual lighting electricity consumption of the building is calculated with smart lamp and standard LED lamp and found as 1747 kWh/year only for daytimes.

ACKNOWLEDGMENT

The authors would like to thank to Computer Engineer Beyza Nur Bora for supporting the research on data measurement.

REFERENCES

[1] Aalto University, Chapter 2: Lighting energy in buildings, 2010.

[2] M. C. Sahin, G. N. Gugul and M. A. Koksal, "Effects of Appliance Standby Electricity Consumption on Turkish Residential Electricity Sector", 7th International Conference on Energy Efficiency in Domestic Appliances and Lighting (EEDAL 13), Coimbra, Portugal, September 2013.

[3] US EIA-Residential Sector, 2018. Available: https://www.eia.gov/todayinenergy/detail.php?id=36412.

[4] S. Onaygil, Aydınlatma Tekniği, Verimlilik, Planlama ve Yönetim, Aydınlatmada Planlama ve Yönetimin Önemi, Gaziantep, 2016.

[5] DIAL, 2020. Available: https://www.dial.de/en/dialux/.

[6] Relux, 2020. Available: https://reluxnet.relux.com/en/.

[7] A. V. d. Silva, A. O. Godinho, C. I. F. Agreira and M. M. T. Valdez, An educational approach to a Lighting Design Simulation using DIALux evo Software, IEEE 51st International Universities Power Engineering Conference, Coimbra, 2016.

[8] K. R. Wagiman and M. N. Abdullah, "Intelligent Lighting Control System for Energy Savings in Office Building", Indonesian Journal of Electrical Engineering and Computer Science, vol. 11, no. 1, pp. 195-202, 2018.

[9] E. Erkin, "Ofis Binaları İçin Aydınlatma Enerjisi Tasarruf Potansiyelleri Hesaplama Amaçlı Bir Yöntem Önerisi", İstanbul Teknik Üniversitesi-Enerji Enstitüsü, İstanbul, 2012.

[10] H. Aydogan and M. F. Özsoy, "Sayısal Aydınlatma Analizi İçin Bir Yazılım Geliştirilmesi", Eğitim ve Öğretim Araştırmaları Dergisi, vol. 6, no. 2, pp. 316-321, 2017.

[11] S. R. Ali, L. Mahjdoubi and A. Khan, "A Study of Different Building Energy Lighting", Advances in Energy and Power, vol. 3, no. 4, pp. 91-95, 2015.

[12] A. Karamouzi, D. Papalexopoulos, A. Stavridou, S. Tzimopoulou and T. Varoudis, "D.L.D. Dynamic Lighting Design-Parametric interactive lighting software in urban public space", IEEE 9th International Conference on Intelligent Environments , 2013.

[13] J. F. D. Paz, J. Bajo, S. Rodríguez, G. Villarrubia and J. M. Corchado, "Intelligent system for lighting control in smart cities", Information Sciences, vol. 372, pp. 241-255, 2016.

[14] S. A. Mahajan and S. D. Markande, "Design of intelligent system for indoor lighting", IEEE International Conference on Computing Communication Control and automation (ICCUBEA), Pune, 2016.

[15] J. Zhou, H. Guo, B. Bai, M. Duan and C. Lin, "Power-white LED dimming detection system based on NC PWM", 11th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID), Xiamen, 2017.

[16] T. Wanga, T. Chen, Y. Hu, X. Zhou and N. Song, "Design of intelligent LED lighting systems based on STC89C52 microcomputer", Optik, vol. 158, pp. 1095–1102, 2018.

[17] N. Savvinova, "Development of a Smart Lamp With a Motion Sensor", eLIBRARY ID: 38585280, pp. 57-60, 2019.

[18] EMO, Lighting, 2020. Available: http://www.emo.org.tr/genel/bizden_detay.php?kod=48544&tipi=34&sube=0&harf=A.

[19] D. Crawley and L. Lawrie, Climate.OneBuilding.Org, 2020. Available: http://climate.onebuilding.org/default.html.

[20] P. M. Pattison, M. Hansen and J. Y. Tsao, "Respond to LED Lighting Efficacy: Status and Directions", U.S. Department of Energy Solid State Lighting Program, Washington, DC, 2017.

An Accuracy Improvement of the Neuromorphic Functional Models by Using the Parallel ANN Architecture

Sergey Mosin
Institute of Computational Mathematics and Information Technologies
Kazan Federal University (KFU)
Kazan, Russian Federation
smosin@ieee.org

***Abstract*—Enhancement of the up-to-date computing systems in performance and memory capacity stimulates development of new mathematical models and methods for numerical simulation. Machine learning methods are widely used nowadays in the electronic design automation. New mathematical entities are focused onto increasing the design quality and reducing a time cost. A method of constructing the neuromorphic functional models (NFM) for analog components and functional blocks is proposed. An approach to improvement of the NFM accuracy by partitioning the domain of definition for output characteristics according to the threshold coefficient and using the parallel artificial neural network (ANN) architecture is offered. The automated synthesis route of the NFM is represented. The results of experimental study for semiconductor diode and the voltage rectifier circuit are demonstrated. The accuracy increasing of the synthesized NFM and circuit simulation results shown high efficiency of the proposed method.**

Keywords—machine learning, neuromorphic functional models, analog components and functional blocks, design automation

I. Introduction

The integrated technologies development directly affected a significant increase in the functionality of modern microprocessors and computing systems in general. The increase in CPU/GPU performance and memory capacity allows solving the complex computational problems in an acceptable time with high accuracy. One of the sources of highly loaded computing problems is design automation in various sectors of the economy and in microelectronics in particular. The development of CAD tools for microelectronics allows us to design new microprocessor devices. Thus, we can state that the developments of CAD tools and computer systems are interrelated processes influencing each other.

Mathematical models and mathematical methods are the two main entities of mathematical support in the state-of-the art CAD tools. The growth of the functional possibilities of computing systems stimulates the further development of new mathematical models and methods that take into account the peculiarities of the computing systems architecture, including the strategy of parallel processing [1]. Particular attention is now paid to mathematical models of components and functional blocks that take into account the features of integrated technology and ensure the required design quality [2–6], as well as design methods and technologies [7–10] that provide the reliability and high quality of the developed devices within the framework of up-to-date design methodologies like Design-for-Testability (DFT) [11–14], Design-for-Manufacturing (DFM) [15], etc.

The development of integrated technologies has influenced the emergence of new architectures and concepts for their development like Application-Specific Integrated Circuits (ASIC), System-on-a-Chip (SoC), Network-on-a-Chip (NoC), embedded-systems, etc. At the same time, the main technical and economic criteria for designing became a reduction in development time and time to market, as well as a costs reduction.

The methods of artificial intelligence (AI) and machine learning (ML) have been actively used recently in the electronic design automation at the level of acceleration of computations in the hardware [16] and software at solving systems of equations [17–18], reliability assessment [19–21], evolutionary optimization [22], building models of devices and systems [23–24], etc.

A method for constructing the functional models of the components and functional blocks of analog and mixed-signal circuits for the circuit level design, based on the use of artificial neural networks, is proposed. An approach to increase the accuracy of the mathematical model by pre-processing the initial data and using the parallel architecture of the ANN is also offered.

The paper is organized as the following. Section II describes the mathematical statement of the problem on constructing a neuromorphic functional model (NFM). The automated synthesis route of the NFM is proposed in Section III. The experimental results are presented in Section IV, while the next section contains the conclusion and future work outlines.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

II. Mathematical Statement of the Problem on Constructing a Neuromorphic Functional Model

Artificial neural networks (ANNs) are actively used in solving the problems of clusterization, classification and regression. In the first case, unsupervised training methods are used, and in the second and third cases, the supervised training methods are applied.

The ability of ANN to approximate functional dependencies underlies the construction of neuromorphic functional models (NFMs) of analog components and functional blocks. Models of analog components and functional blocks used in circuit design are based on Ohm and Kirchhoff laws and reflect the relationship between the current flowing through the component and the applied voltage, taking into account many external and internal parameters

$$\mathbf{Y}_{out} = f\left(\mathbf{X}_{in}\right) \tag{1}$$

where $\mathbf{X}_{in}$ is a vector of the input values, $\mathbf{Y}_{out}$ is an associative vector of the output values.

The current flowing through the component ($I \in \mathbf{Y}_{out}$) is considered at the NFM as the output characteristic for the one-terminal components (Fig. 1), and the applied voltage to the component is the input characteristic ($V \in \mathbf{X}_{in}$).

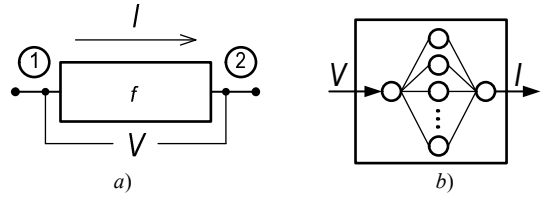


Fig 1. One-terminal component: schematic (*a*), NFM (*b*)

The input and output currents ($I_{out}, I_{in} \in \mathbf{Y}_{out}$) are considered at the NFM as the output characteristics for the two-terminal components (Fig. 2), and the input current ($I_{in} \in \mathbf{X}_{in}$), the input voltage ($V_{in} \in \mathbf{X}_{in}$) and the output voltage ($V_{out} \in \mathbf{X}_{in}$) are considered as the input characteristics.

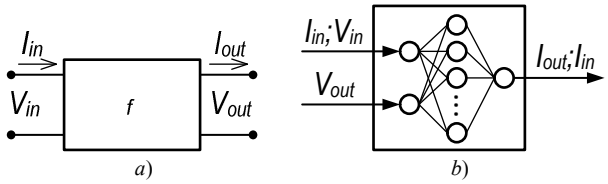


Fig. 2. Two-terminal component: schematic (*a*), NFM (*b*)

In the general case, with an increase in the range of variation of the input values, a significant increase in the values of the output characteristic is observed. The domain of definition of the output characteristic may include values that are several orders of magnitude different from each other, in the range of values of the input characteristic

$$f : \mathbf{X} \to \mathbf{Y}, \tag{2}$$

$$y_{max} = \max\left(\mathbf{Y}\right),\ y_{min} = \min\left(\mathbf{Y}\right),$$

$$y_{max} \gg y_{min}.$$

A significant dispersion in the values of the output characteristic affects the decrease in the accuracy of the approximation f (2) using an ANN. Improving the accuracy of approximation and the quality of ANN training is achieved by pre-processing the source data, which is represented by an array of tuples in the following form

$$\mathbf{M} = \left\{ m_n = \left\langle x_1^{(n)}, ..., x_{N_x}^{(n)}, y_1^{(n)}, ..., y_{N_y}^{(n)} \right\rangle \right\}, \tag{3}$$

$$x_i \in \mathbf{X}_{in},\ y_k \in \mathbf{Y}_{out},\ i = 1..N_x,\ k = 1..N_y,$$

where N_x is the number of the input parameters of the model, N_y is the number of the output parameters of the model, N_s is the number of discrete values of the functional dependence (1) in N_x-dimensional space of changes input values x_i

$$S_i^{(n)} \le x_i^{(n)} \le E_i^{(n)},\ \forall i = 1..N_x,$$
$$S_i^{(n)} = \min\left(x_i^{(n)}\right),\ E_i^{(n)} = \max\left(x_i^{(n)}\right).$$

Normalization, for example, linear transformation to a unit scale, usually acts as the raw data preprocessing

$$\tilde{y}_k^{(n)} = \frac{y_k^{(n)} - y_{k,min}^{(n)}}{y_{k,max}^{(n)} - y_{k,min}^{(n)}}, \tag{4}$$

$$y_{k,min}^{(n)} = \min\left(y_k^{(n)}\right),\ y_{k,max}^{(n)} = \max\left(y_k^{(n)}\right),$$

$$k = 1..N_y,\ n = 1..|\mathbf{M}|$$

In the case of $y_{k,min}^{(n)} \ll y_{k,max}^{(n)}$, for $\forall y_j^{(n)} \approx y_{k,min}^{(n)}$ there is a loss of a significant part of the number at the ANN training.

A logarithmic transformation of the range of definition with the base P can be used also as a preprocessing

$$\mathbf{Y}_L = f_{log} : \mathbf{Y} \mapsto \log_P\left(\mathbf{Y}\right). \tag{5}$$

Further, an exponential (power) restoration of the original is carried out during the post-processing

$$\mathbf{Y}_O = f_{exp} : \mathbf{Y}_L \mapsto P^{\wedge}\left(\mathbf{Y}_L\right). \qquad (6)$$

In this case, minor ANN training errors lead to significant errors in restoring the original numerical value of the output characteristic due to the exponential nature of the dependence.

As an alternative, the partitioning the domain of definition (2) into subdomain with the same order of significance of the elements included in them and carrying out independent ANNs training for each formed subdomain are proposed for training the neuromorphic functional model to ensure the approximation accuracy

$$\mathbf{M} = \mathbf{M}_1 \cup \mathbf{M}_2 \cup \ldots \cup \mathbf{M}_k \cup \ldots \cup \mathbf{M}_G, \qquad (7)$$

$$\mathbf{M}_k = \left\{ m_{k1}, m_{k2}, \ldots, m_{kl}, \ldots, m_{kp_k} \right\},$$

$$m_{kl} = \left\langle x_1^{(kl)}, \ldots, x_{N_x}^{(kl)}, y_1^{(kl)}, \ldots, y_{N_y}^{(kl)} \right\rangle,$$

$$\theta_{Lo} \le \frac{y_j^{(k1)}}{y_j^{(kp_k)}} \le \theta_{Hi}, \ \forall k = 1..G, j = 1..N_y,$$

$$\theta_{Hi} < \frac{y_j^{(k1)}}{y_j^{(Q1)}} \text{ or } \theta_{Lo} > \frac{y_j^{(k1)}}{y_j^{(Q1)}}, \ \forall k = 1..G, \forall Q = 1..G,$$

$$k \ne Q, j = 1..N_y, \ \theta_{Hi} = \theta, \theta_{Lo} = 1/\theta_{Hi},$$

where θ is the threshold partitioning coefficient, p_k is the cardinality of k-th subdomain.

The resulting approximator is represented by the parallel ANN architecture, combining the G ANNs for all subdomains (7). The parallel ANN architecture is presented in Figure 3.

III. The Automated Synthesis Route of the NFM

The design flow of an automated synthesis of the NFM can be described by the following sequence of steps (Fig. 4).

1. Generating the initial data, which can be performed by two ways, firstly, based on the results of modeling the analytical dependence of the current on the voltage (model-based) or, secondly, based on the results of measuring the characteristics during physical testing of a component (data-driven). An array of tuples is used for structuring and representation of the raw data according to (3).

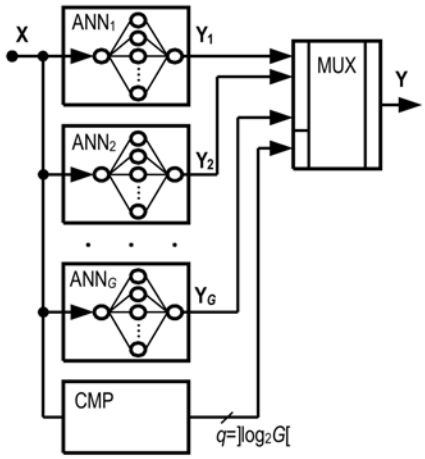


Fig. 3. The parallel ANN architecture

2. Partitioning the raw data according to specified threshold coefficient θ. The complete dataset is split on several subsets (7) by the criterion of difference the values of the output characteristics no more than in θ times inside one subdomain.

3. Selecting the ANN$_i$ architecture ($i = 1..G$) is focused on determining the number of layers, the number of neurons in each layer and the type of activation function. A two-layer perceptron is used as a base of the NFM. The number of input parameters of the model determines the number of neurons in the input layer N_x. The number of model output parameters determines the number of neurons in the output layer N_y. The number of neurons of the hidden layer (N) of the two-layer perceptron is estimated taking into account the cardinality of i-th training subset, as well as dimensions of the input (**X**) and output (**Y**) sequences.

4. Generating the training ($\mathbf{M}_i^{trn}$) and the testing ($\mathbf{M}_i^{tst}$) subsets by a uniform sampling from the subset of raw data $\mathbf{M}_i$

$$\mathbf{M}_i^{trn} \in \mathbf{M}_i, \ \mathbf{M}_i^{tst} \in \mathbf{M}_i, \ \mathbf{M}_i^{trn} \cap \mathbf{M}_i^{tst} = \varnothing,$$

$$k_i^{trn} = \left|\mathbf{M}_i^{trn}\right|, \ k_i^{tst} = \left|\mathbf{M}_i^{tst}\right|, \ k_i^{trn} < k_i^{tst},$$

where k_i^{trn} is the number of elements in the i-th training subset, k_i^{tst} is the number of elements in the i-th testing subset.

5. Training of the ANN$_i$ is executed using the corresponding subset $\mathbf{M}_i^{trn}$. The training process is stopped when either the training error is became less or equal to the threshold value, or when the number of executed iterations exceed the maximum available value. The quality of the ANN$_i$ training is tested using the subset $\mathbf{M}_i^{tst}$.

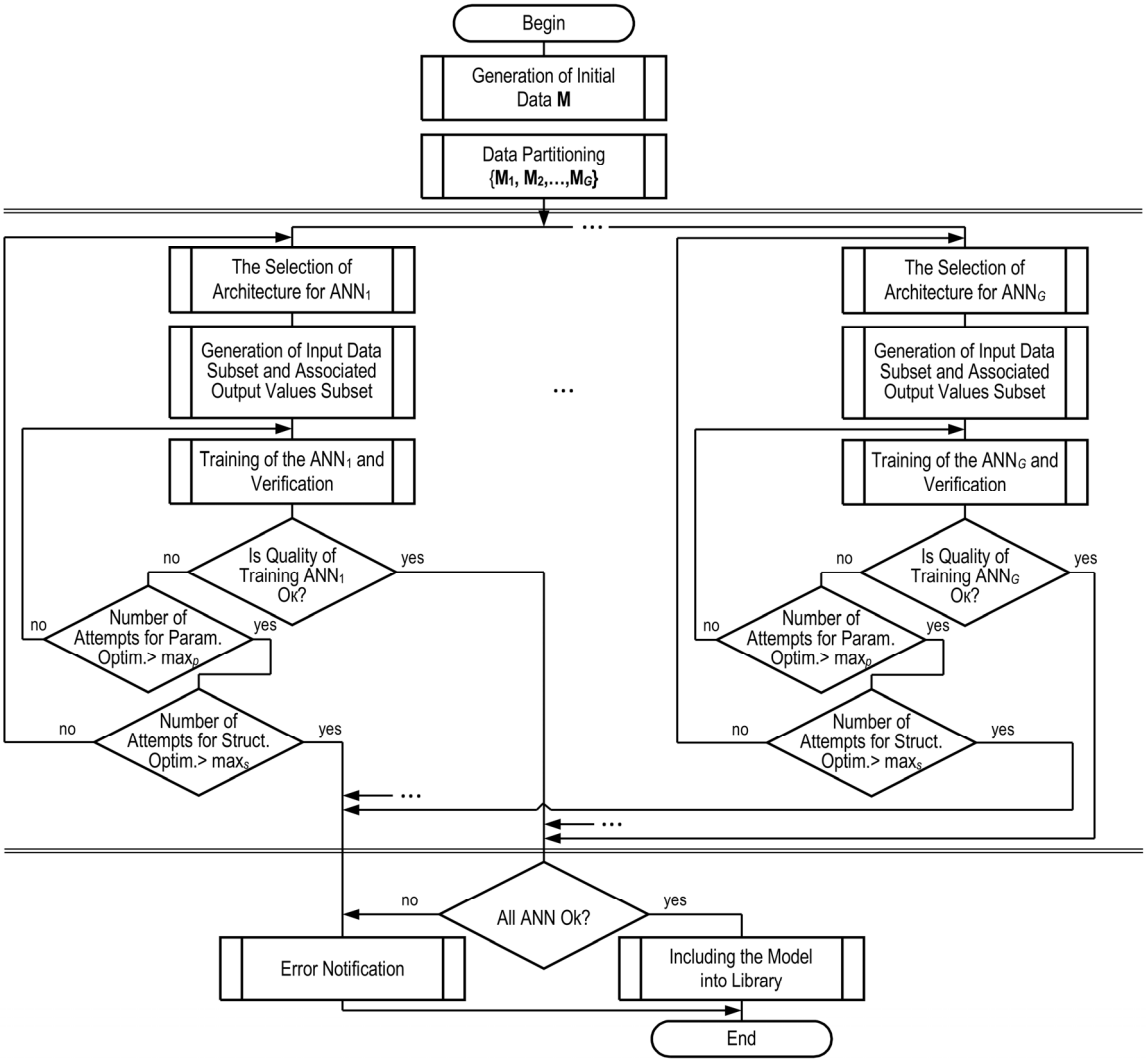


Fig. 4. The route of automated synthesis of the NFM

A cycle of the parametric synthesis is initiated if the required quality of the ANN_i training has not been achieved. In this case the ANN_i architecture selected on the step 3 is re-trained with random re-assigning the initial conditions.

A cycle of the structural synthesis is initiated if the required quality of the ANN_i training cannot be achieved during the limited number of parametric synthesis attempts ($\max_p$). This cycle dealt with a return to the step 3 and making modifications in the ANN_i architecture like a changing the number of neurons in the hidden layer, increasing the number of intermediate layers, etc.

The process is stopped with generation of the corresponding notification if after $\max_s$ attempts the structural synthesis could not provide required quality of the ANN_i training.

Steps 3-5 are executed independently for each ANN_i but in the parallel way for all G subdomains simultaneously. The parallel execution of these steps ensures the effective use of the up-to-date computing systems with reducing the time cost on the model synthesis.

6. The successfully trained G ANNs have represented the complete model, which is stored in the library for the further application during describing and simulating the electronic circuits.

IV. Experimental Results

A. Construction of the Neuromorphic Functional Model

The synthesis of a neuromorphic functional model is demonstrated using the semiconductor diode D1N4934 as a case study. The raw data has been generated during simulation of the diode structural model in the Cadence CAD tools. The Volt-Ampere Characteristic (VAC) represents the initial data for the NFM synthesis. The VAC reflects the dependence of the current flowing through the diode on the applied voltage from the range –6 V up to +6 V with step 0.01 V and the corresponding

current is changed from –1.573e–07 A to 1.239e+02 A. In result the raw data **M** consists of 12 001 tuples $m_n = \langle V_n, I_n \rangle$, where V_n is the effective applied voltage to the diode and I_n is the corresponding current flowing through the diode.

The partition of the initial data is performed for two values of the threshold partitioning coefficients θ = 5 and θ = 10. The number of subdomains is equal to 15 in the first case and 11 in the second case.

A two-layer perceptron with one neuron on the input, one neuron on the output and *N* neurons in the hidden layer was used for implementation of the NFM as a basic architecture for each subdomain dataset. The number of neurons in the hidden layer is estimated taking into account actual number of samples in the corresponding subdomain.

The comparative results of the approximation quality for the trained neuromorphic functional models are represented in Table I. The approximation quality is estimated for the NFM synthesized on the set of initial data without partitioning (G = 1) and for both cases of partitioning (G = 15 and G = 11).

TABLE I. COMPARATIVE RESULTS OF THE APPROXIMATION QUALITY

Threshold partitioning coefficient (θ)	Number of Subdomains (*G*)	The average approximation value	Root mean square error
1	1	0.1391e+01	1.5076e+04
5	15	0.5890e–03	5.6016e+03
10	11	0.5830e–03	5.3245e+03

The tool of mathematical and engineering calculation MATLAB and the computing system with processor Intel®Core™ i7-4770 CPU @3.4GHz and RAM 8GB were used for training the NFM. Obtained results demonstrate an essential increasing the quality of approximation after use of the proposed approach, the trained NFMs based on the partitioning and parallel architecture provide improvement of the average approximation value in more 2 361 times and RMSE in about 2.96 times.

B. Numerical Simulation of the Circuit with Using the Neuromorphic Functional Model

The NFM trained without partitioning of the initial data and the NFM trained at partitioning with θ = 10 were used for description and numerical simulation of the two-wave bridge rectifier as a testbench (Fig. 5).

The results of the numerical simulation of the circuit's mathematical model based on the NFMs ($NFM_{\Theta=1}$ without partitioning and $NFM_{\Theta=10}$ at partitioning with threshold coefficient 10) were compared with the results of simulating the corresponding rectifier's circuit in the Cadence CAD tools (Table II).

The numerical simulation of the two-wave bridge rectifier using the NFM for the diodes D_1–D_4 was executed for two types of the circuit analysis. The first type is DC-analysis or analysis in the static mode with changing the input voltage V_{in} in the range –6V up to +6V with step 0.1V. The second type is the transient analysis or analysis in the time domain with changes of the input voltage V_{in} according to sine-wave low with amplitude 220V and frequency 5 kHz during three periods.

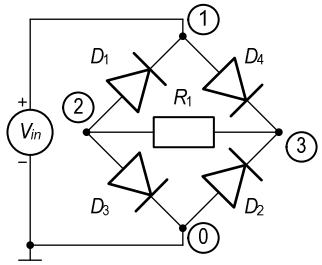


Fig. 5. The voltage two-wave bridge rectifier

TABLE II. THE CIRCUIT SIMULATION ERRORS

Error	DC		Tran	
	$NFM_{\Theta=1}$	$NFM_{\Theta=10}$	$NFM_{\Theta=1}$	$NFM_{\Theta=10}$
Max_Rel, %	0.199e+01	0.426e+00	0.11e+00	0.231e–01
Avg_Rel, %	0.199e+01	0.249e–01	0.94e+00	0.110e–01
Max_Abs, V	0.267e–01	0.104e–03	0.499e+00	0.498e+00
Avg_Abs, V	0.175e–01	0.357e–04	0.125e+00	0.362e–01
RMSE, V	0.149e–01	0.350e–04	0.153e+00	0.101e–00

The combined graphs of the simulating results for the two-wave bridge rectifier based on the $NFM_{\Theta=10}$ in the static mode and in the time domain are presented in Figure 6.

V. CONCLUSION

The proposed approach to synthesis of the neuromorphic functional models based on the partitioning the domain of definition on subdomains and training the parallel ANN architecture has demonstrated high efficiency. The synthesized NFM provide increasing the accuracy of approximating the functional dependence of output characteristics on the input characteristics. The proposed approach ensures decreasing the relative and absolute errors of the circuit simulation in the static mode and the time domain for the considered example in 1.51 up to 490 times. Once trained ANN as the NFM for analog components or functional blocks is stored in the library and can be regularly used for a circuit description and simulation.

The proposed route of automated synthesis of the NFM is realized in the MATLAB software and can be used for analog circuit design in the framework of design flow. The synthesized NFM can be implemented in the hardware realizing artificial neural networks.

ACKNOWLEDGEMENT

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

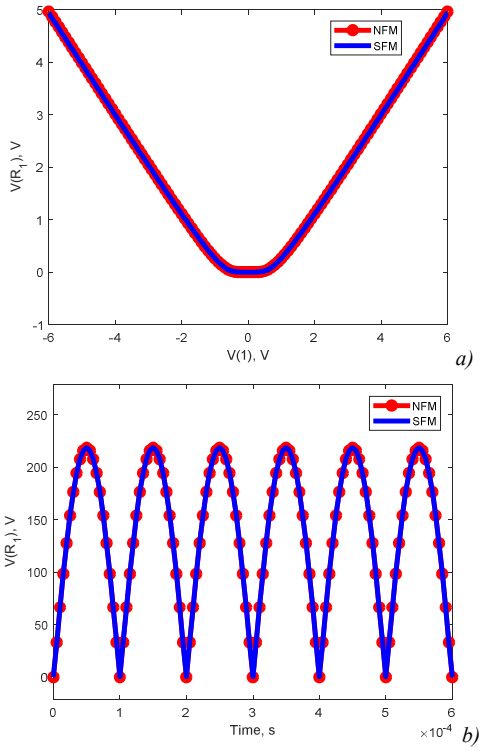


Fig. 6. Combined graphs of output voltage for the two-wave bridge rectifier with application of the partitioned neuromorphic (NFM) and structural (SFM) functional models: in the static mode (*a*), in the time domain (*b*)

REFERENCES

[1] S. Mosin, "The State-of-the-Art Trends in Education Strategy for Sustainable Development of the High Performance Computing Ecosystem," Communications in Computer and Information Science. Springer, Cham, vol. 793, 2017, pp. 494-504.

[2] M. Ho et al., "Architecture and Design Flow for a Highly Efficient Structured ASIC," in IEEE Trans. on Very Large Scale Integration (VLSI) Systems, vol. 21, no. 3, 2013, pp. 424-433.

[3] K. Nepal, S. Hashemi, H. Tann, R. I. Bahar and S. Reda, "Automated High-Level Generation of Low-Power Approximate Computing Circuits," in IEEE Trans. on Emerging Topics in Computing, vol. 7, no. 1, 2019, pp. 18-30.

[4] A.C. Oliveira, P.C.C. de Aguirre, L.C. Severo, A.G. Girardi, "An optimization-based methodology for efficient design of fully differential amplifiers," in Analog Integrated Circuits and Signal Processing, vol. 90, no. 1, 2017, pp. 149-163.

[5] K.O. Petrosyants, I.A. Kharitonov, S.V. Lebedev, L.M. Sambursky, S.O. Safonov, V.G. Stakhin, "Electrical characterization and reliability of submicron SOI CMOS technology in the extended temperature range (to 300 °C)," in Microelectronics Reliability, vol. 79, 2017, pp. 416-425.

[6] H. Zou, Y. Moursy, R. Iskander et al., "A CAD integrated solution of substrate modeling for industrial IC design," in Proceedings of the 2015 IEEE 20th International Mixed-Signals Test Workshop, IMSTW 2015, Paper № 7177885, 2015.

[7] A.S. Adonin, K.O. Petrosyants, D.A. Popov, "Modeling of the submicron MOSFETs characteristics for UTSi technology," in Proc. of SPIE - The International Society for Optical Engineering, 11022, Paper № 110220G, 2019.

[8] A.M. Pilipenko, V.N. Biryukov, N.N. Prokopenko, "A Template Model of Junction Field-Effect Transistors for a Wide Temperature Range," in Proc. of IEEE East-West Design and Test Symposium, EWDTS 2019, Paper № 8884411, 2019.

[9] N.N. Prokopenko, A.R. Gaiduk, A.V. Bugakova, E.V. Ovsepian, "Mathematical analysis of transients of the high-speed buffer amplifier with the complementary composite transistors in nonlinear mode," in Proc. of 23rd International Conference on System Theory, Control and Computing, ICSTCC 2019, Paper № 8885938, 2019, pp. 292-297.

[10] A.L. Stempkovsky, A.D. Ivannikov, "Formal Description of Digital Control System Operation and Its Use in Designing," in Russian Microelectronics, vol. 48, no. 5, 2019, pp. 318-325.

[11] D.V. Efanov, V.V. Sapozhnikov, V.V. Sapozhnikov, "Using Codes with Summation of Weighted Bits to Organize Checking of Combinational Logical Devices," in Automatic Control and Computer Sciences, vol. 53, no. 1, 2019, pp. 1–11.

[12] D.V. Efanov, V.V. Sapozhnikov, V.V. Sapozhnikov, "Sum Codes with Fixed Values of Multiplicities for Detectable Unidirectional and Asymmetrical Errors for Technical Diagnostics of Discrete Systems," in Automation and Remote Control, vol. 80, no. 6, 2019, pp. 1082-1097.

[13] S. Lerner, I. Yilmaz and B. Taskin, "Custard: ASIC Workload-Aware Reliable Design for Multicore IoT Processors," in IEEE Trans. on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 3, 2019, pp. 700-710.

[14] S. Mosin, "Design-for-testability automation of mixed-signal integrated circuits," in Proc. of International System on Chip Conference, Paper № 6749695, 2013, pp. 244-249.

[15] O.V. Dvornikov, N.N. Prokopenko, V.A. Tchekhovski, Y.D. Galkin, A.V. Kunz, A.V. Bugakova, "Test Chip for Identifying Spice-Parameters of Cryogenic BiFET Circuits," in Proc. of European Solid-State Device Research Conference, Paper № 8901773, 2019, pp. 102-105.

[16] Y. Li and Y. Du, "A Novel Software-Defined Convolutional Neural Networks Accelerator," in IEEE Access, vol. 7, 2019, pp. 177922-177931.

[17] C. Michoski, M. Milosavljević, T. Oliver, D.R. Hatch, "Solving differential equations using deep neural networks," in Neurocomputing, vol. 399, 2020, pp. 193-212.

[18] V. Dwivedi, B. Srinivasan, "Physics Informed Extreme Learning Machine (PIELM)–A rapid method for the numerical solution of partial differential equations," in Neurocomputing, vol. 391, 2020, pp. 96-118.

[19] R. Dautov, S. Mosin, "A technique to aggregate classes of analog fault diagnostic data based on association rule mining," in Proc. of International Symposium on Quality Electronic Design, ISQED, 2018, pp. 238-243.

[20] S.G. Mosin, "On the Construction of Neuromorphic Fault Dictionaries for Analog Integrated Circuits," in Russian Microelectronics, vol. 48, no. 5, 2019, pp. 310-317.

[21] S. Mosin, "Machine learning and data mining methods in testing and diagnostics of analog and mixed-signal integrated circuits: Case study," in Communications in Computer and Information Science, vol. 968, 2019, pp. 240-255.

[22] J. Koza, F. Bennett III, D. Andre, and M. Keane, "The design of analogue circuits by means of genetic programming," in Evolutionary Design by Computers, P. J. Bentley, Ed. John Wiley&Son, 1999, ch. 16, pp. 365–385.

[23] Q. Chen and G. Chen, "Artificial neural network compact model for TFTs," in Proc. of 7th International Conference on Computer Aided Design for Thin-Film Transistor Technologies (CAD-TFT), Beijing, 2016, pp. 1-1.

[24] E.B. Solovyeva, "Behavioural nonlinear system models specified by various types of neural networks," in Journal of Physics: Conference Series, vol. 1015, no. 3, 2018, Paper № 032139.

Classification of Errors in Ternary Code Vectors from the Standpoint of Their Use in the Synthesis of Self-Checking Digital Systems

Dmitry Efanov,
DSc, Professor at Higher School of Transport,
Institute of Mechanical Engineering, Materials and Transport,
Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation
TrES-4b@yandex.ru

Abstract—The author of the article analyzes the errors that occur in ternary code vectors from the standpoint of their use in the synthesis of self-checking digital systems. The article discusses the issues of identifying the typical types of errors that occur in binary and ternary code vectors. Classifications of errors are formed, and also definitions of the main types of errors are given and their key features are noted. It is shown that errors in ternary code vectors are much more diverse than errors in binary code vectors, which is due to the number system and the number of signals used to represent numbers. The article presents the main classes of redundant codes focused on error detection. The author gives an example of ternary redundant codes belonging to certain classes. The article highlights the prospects for their application in the construction of ternary self-checking digital devices, as well as technical means of their diagnostics.

Keywords—digital systems; binary code vectors; ternary code vectors; errors in code vectors; classification of errors in code vectors; codes with error detection; undetectable error.

I. Introduction

Despite the widespread use of binary logic as the basis of digital technology, engineers and scientists around the world have a lively interest in ternary logic and ternary technology [1 – 4]. There have been attempts to implement devices operating in ternary logic, both based on traditional binary logic devices and using ternary logic elements [5 – 9]. The use of ternary logic in the development of quantum computers is also discussed, and the use of cutrites instead of qubits can significantly reduce the number of quantum gates [10]. An important issue is the study of methods for constructing devices and automation systems endowed with the property of fault detection, which is associated with the use of error-tolerant ternary codes [11].

In the development of self-checking binary devices of automation and computer technology, error-tolerant coding methods are widely used [12, 13]. Coding is used at various levels of discrete device and system architecture. For example, when synthesizing devices at the level of models of finite automata, states are often encoded with a certain redundant code, which allows detecting faults in the operation of devices, as well as parrying their manifestations and ensuring the regular operation of devices [14]. In addition, redundant coding is used in the synthesis of diagnostic tools [15], as well as in data transfer between nodes of automation systems [16]. Similar principles can be applied in the development of self-checking digital devices operating in ternary logic.

The choice of a code with certain characteristics for detecting errors of various types and multiplicities is fundamental in the process of developing self-checking automation devices based on error-tolerant codes [17]. From this point of view, errors in binary code vectors are usually classified into monotonous (unidirectional), symmetrical and asymmetrical, and there are also special classes of redundant codes that have the ability to detect any errors of a certain type and any errors of a certain type up to their specific multiplicity [18, 19].

This article reveals the features of error classification in ternary code vectors in comparison with the features of error classification in binary code vectors. Special types of errors are identified, and classes of ternary redundant codes are introduced, the use of which makes it possible to synthesize self-checking digital devices that function in ternary logic.

II. Errors in Binary and Ternary Vectors and Their Classification

A. Errors in binary code vectors

Let's consider the features of errors that occur in binary and ternary code vectors.

Let's draw an analogy between a code vector and a device with a certain number of outputs: each output will correspond to a certain bit of the code vector, and the value of this bit will be formed by calculating the output function. The correct code vector is formed as a result of the regular operation of the device. The occurrence of a fault in the structure of the device leads to the detection of an error at its output under certain conditions (if the error observability conditions are met). The error can be transmitted to one or several outputs, as well as in different ways: either change its appearance, or save it. For example, if there are inversions on the way to the device's output from the place where the fault occurred, the error will change its type when transmitting to the output. Thus, an error at the output of any element of the digital device structure can lead to an error in the code vector, and these errors can be different in the composition of the distorted values. For example, a device that operates in binary logic, and does not have inverters, will have the property of transmitting the type of error on each line to its outputs. This feature of the device is related to the fact that the functions it implements will be monoto-

978-1-7281-9899-6/20/$31.00 ©2020

nous, and any distortion on the circuit line will only lead to monotonous manifestations of errors. In other words, the property of monotonicity of functions is directly related to the property of monotonicity of the error that occurs in the code vector. The class of monotonous functions is one of the five main classes of the Boolean functions [20]: the class T_0 is the zero-preserving functions; the class T_1 is the one-preserving functions; class M is the monotonous functions; class L is the linear functions; class S is the self-dual functions.

In general, the basic classes of the Boolean functions are widely used in the development of technical diagnosis methods. For example, the linear parity function is widely used in the control of calculations [21 – 23], codes with detection of any monotonous errors are used in the concurrent error-detection (CED) systems of logic circuits [24], as well as control for the belonging of functions to the class of self-dual [25]. The type of errors at the outputs of diagnostic objects has the key importance in the organization of diagnostic support. Often, as noted above, an analogy is drawn between the outputs of automation devices and code vectors, and the error at the outputs of the devices is compared with the error in the code vector [26].

Definition 1. *An error* in a code vector is a set of distortions of its bits.

Let the number of bits of the code vector generated at the outputs of a certain diagnostic object be m. Then the error in the code vector can be associated with a distortion from 1 to m bits.

Definition 2. The number of bits that are distorted when an error occurs is called *the d multiplicity of the error*.

Errors (ERR) in code vectors can be single (one-time) (SIN) and multiple (MULT).

The total number of errors in binary code vectors is determined by the doubled number of transitions of each of 2^m code vectors to each:

$$N_m^{BIN} = 2C_{2^m}^2 = 2 \cdot \frac{2^m}{2!\left(2^m-2\right)!} = \frac{2 \cdot 2^m\left(2^m-1\right)\left(2^m-2\right)!}{2!\left(2^m-2\right)!} = 2^m\left(2^m-1\right). \quad (1)$$

For example, three-digit binary code vectors are distorted by $N_3^{BIN} = 2C_{2^3}^2 = 2^3\left(2^3-1\right) = 8 \cdot 7 = 56$ variants.

On the other hand, the total number of undetectable errors in binary code vectors is calculated as the sum of all errors of each multiplicity $d \in \{1,2,...,m\}$:

$$N_m^{BIN} = \sum_{d=1}^{m} 2^m \cdot 1^d \cdot C_m^d = \sum_{d=1}^{m} 2^m C_m^d = 2^m \sum_{d=1}^{m} C_m^d, \quad (2)$$

where the 2^m cofactor determines the total number of binary code vectors, the 1^d cofactor characterizes the number of distortion variants of d bits, and the C_m^d cofactor is the number of distortions of multiplicity d out of m bits.

For example, for a variant with m=3, formula (2) gives the following result:

$$N_3^{BIN} = 2^3 \sum_{d=1}^{m=3} C_3^d = 2^3 \cdot \left(C_3^1 + C_3^2 + C_3^3\right) = 8 \cdot \left(3+3+1\right) = 8 \cdot 7 = 56.$$

In two-valued logic, each bit of a code vector can contain two types of distortion: 0→1 or 1→0. The type of error that occurs is determined by the set of distortions of the various bits.

Definition 3. An error is called *monotonous (unidirectional)* if the binary code vector contains only distortions of the type 0→1 or only of the type 1→0, when an error occurs.

Definition 4. An error is called *non-monotonous (multidirectional)* if there are various types of distortions in the binary code vector when an error occurs.

Single (SIN), monotonous (MON) and non-monotonous (NMON) errors form the full set of errors in binary code vectors. Non-monotonous errors are usually divided into two unequal classes of symmetrical (SYM) and asymmetrical (ASYM) errors.

Definition 5. An error is called *symmetrical* if it occurs in a binary code vector with the same number of distortions of type 0→1 and type 1→0.

Definition 6. An error is called *asymmetrical* if it occurs in a binary code vector with the different number of distortions of type 0→1 and type 1→0.

The Fig. 1 gives the examples of various types of errors in binary code vectors, and Fig. 2 gives their full classification. It should be noted that the monotonous error can have a multiplicity of $d \geq 2$, the symmetrical error always has an even value of multiplicity, and the asymmetric error has a $d \geq 3$ value of multiplicity.

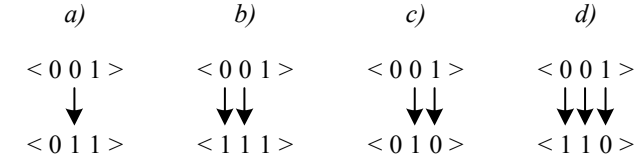
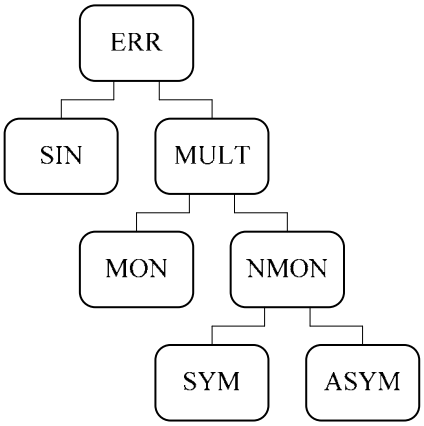


Fig. 1. The examples of various types of errors in the binary vectors: a) single (SIN); b) monotonous (MON); c) symmetrical (SYM); d) asymmetrical (ASYM).

The Table 1 for example gives an error characteristic for binary code vectors with a length m=3, where the number of errors of various types and different multiplicities is indicated. In addition, the last column of the table shows the relative indicator τ, which shows how much the number of errors of a particular type takes up from the total number of errors in code vectors. The number of monotonous and symmetrical errors is approximately the same, while the number of asymmetrical errors is approximately two times smaller. The error distributions by types and multiplicities and the values of indicators τ differ for different lengths of code vectors.



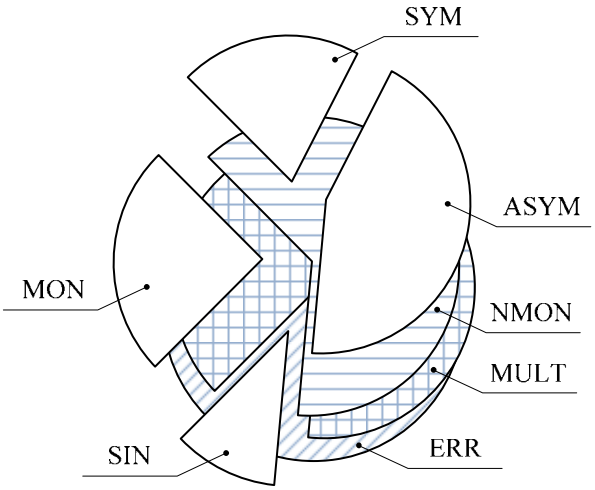


Fig. 2. The classification of errors in binary vectors.

TABLE I. THE DISTRIBUTION OF ERRORS BY TYPES AND MULTIPLICITIES IN BINARY CODE VECTORS WITH THE LENGTH M=3

Error type	Error multiplicity			Total	τ, %
	1	2	3		
SIN	24	–	–	24	42.857
MON	–	12	2	14	25
SYM	–	12	0	12	21.429
ASYM	–	–	6	6	10.714
Total	24	24	8	56	100

B. Errors in ternary code vectors

Now focus on the ternary logic code vectors.

The mathematics of ternary logic uses a large number of values (let's denote them as 0, 1, and 2). Accordingly, the number of basic classes of ternary logic functions is greater than of binary logic: there are 18 basic classes of ternary logic functions [20]. The classes of monotonous functions have a special significance among them. The M_1 class includes monotonous functions for which the order $0 < 1 < 2$ is used when comparing arguments. The M_2 class includes monotonous functions for which the order $1 < 2 < 0$ is used when comparing arguments. The M_3 class includes monotonous functions for which the order $2 < 0 < 1$ is used when comparing arguments.

Based on the classification of errors in binary code vectors, we classify errors in ternary code vectors, putting in it the features of combinations of distortions of 0→1, 1→0, 0→2, 2→0, 1→2, 2→1 types.

The total number of errors in ternary code vectors is determined by the doubled number of transitions of each of 3^m code vectors to each:

$$N_m^{TER} = 2C_{3^m}^2 = 2 \cdot \frac{3^m}{2!\left(3^m-2\right)!} = \frac{2 \cdot 3^m \left(3^m-1\right)\left(3^m-2\right)!}{2!\left(3^m-2\right)!} = 3^m\left(3^m-1\right). \quad (3)$$

For example, three-bit binary code vectors are distorted by $N_3^{TER} = 2C_{3^3}^2 = 3^3\left(3^3-1\right) = 27 \cdot 26 = 702$ variants. This number is 12.54 times greater than the number of errors in binary vectors. As the code vector length increases, the difference between the number of errors in ternary and binary vectors increases rapidly (Table 2). In the limit for $m \to \infty$ the difference is determined by the following value:

$$\lim_{m\to\infty} \delta_m = \lim_{m\to\infty} \frac{N_m^{TER}}{N_m^{BIN}} = \lim_{m\to\infty} \frac{3^m\left(3^m-1\right)}{2^m\left(2^m-1\right)} = \lim_{m\to\infty}\left(\frac{3}{2}\right)^m \lim_{m\to\infty} \frac{3^m-1}{2^m-1} = \infty. \quad (4)$$

TABLE II. THE RELATION BETWEEN THE NUMBERS OF ERRORS IN BINARY AND TERNARY CODE VECTORS

m	N_m^{BIN}	N_m^{TER}	$\delta_m = \frac{N_m^{TER}}{N_m^{BIN}}$
2	12	72	6
3	56	702	12.54
4	240	6480	27
5	992	58806	59.28
6	4032	530712	131.63
7	16256	4780782	294.09
8	65280	43040160	659.32
9	261632	387400806	1480.71
10	1047552	3486725352	3328.45
…	…	…	…
20	$1.09951 \cdot 10^{12}$	$1.21577 \cdot 10^{19}$	11057343
…	…	…	…
50	$1.26765 \cdot 10^{30}$	$5.15378 \cdot 10^{47}$	$4.066 \cdot 10^{17}$
…	…	…	…
100	$1.60694 \cdot 10^{60}$	$2.65614 \cdot 10^{95}$	$1.653 \cdot 10^{35}$

The number of undetectable errors in ternary code vectors can also be determined as the sum of all errors with each multiplicity $d \in \{1,2,...,m\}$:

$$N_m^{TER} = \sum_{d=1}^{m} 3^m 2^d C_m^d = 3^m \sum_{d=1}^{m} 2^d C_m^d, \quad (5)$$

where the 3^m cofactor determines the total number of ternary code vectors, the 2^d cofactor characterizes the number of variants of distortions of *d* bits (in contrast to binary bits, each ternary bit can be distorted by two variants), the C_m^d cofactor is the number of distortions of multiplicity *d* of *m* bits.

For an example with *m*=3, formula (5) gives the following result:

$$N_3^{TER} = 3^3 \sum_{d=1}^{m=3} 2^d C_3^d = 3^3 \left(2^1 C_3^1 + 2^2 C_3^2 + 2^3 C_3^3\right) =$$
$$= 27 \cdot (6 + 12 + 6) = 27 \cdot 26 = 702.$$

We also note that formula (5) can be generalized for code vectors of *q*-valued logic:

$$N_m^q = q^m \sum_{d=1}^{m} (q-1)^d C_m^d. \quad (6)$$

In addition to the above, errors in ternary code vectors are much more diverse than errors in binary code vectors.

Like errors in binary vectors, errors in ternary vectors are divided into single (SIN) and multiple (MULT). Multiple errors are divided into two classes: monotonous (MON) and non-monotonous errors (NMON).

Definition 7. An error in the ternary code vector is *monotonous* if, when it occurs, the priority of the values defined in each of the classes of monotonous functions M_1, M_2 and M_3 is preserved.

Monotonous errors are divided into unidirectional (UNI) and bidirectional (BIDI).

Definition 8. A monotonous error in a ternary code vector is *unidirectional* if, when it occurs, all distortions occur only in the values 0, or only in the values 1, or only in the values 2.

It should be noted that monotonous unidirectional errors can be divided into two classes: double-side (DS) and triple-side (TS) errors.

Definition 9. A monotonous unidirectional error in a ternary code vector is called a *double-side* error if the same bit values are distorted when it occurs.

Definition 10. A monotonous unidirectional error in a ternary code vector is called a *triple-side* error if the different bit values are distorted when it occurs.

Definition 11. A monotonous error in a ternary code vector is *bidirectional* if, when it occurs, the priority of values defined in each of the classes of monotonous functions M_1, M_2 and M_3 is preserved, and distortions of all value variants into all value variants are achievable.

Definition 12. An error in the ternary code vector is *non-monotonous* if any kind of distortion is possible when it occurs.

Non-monotonous errors are divided into compositional (COMP) and asymmetric (ASYM) errors.

Definition 13. A non-monotonous error in a ternary code vector is *compositional* if the number 0, 1, and 2 is preserved when it occurs.

The compositional error in the ternary code vector is an analogue of the symmetrical error in the binary code vector.

Definition 14. A non-monotonous error in the ternary code vector is *asymmetrical* if, when it occurs, the numbers 0, 1, and 2 are not saved.

The examples of various types of errors in ternary code vectors are shown in the Fig. 3. And errors classification is shown in the Fig 4.

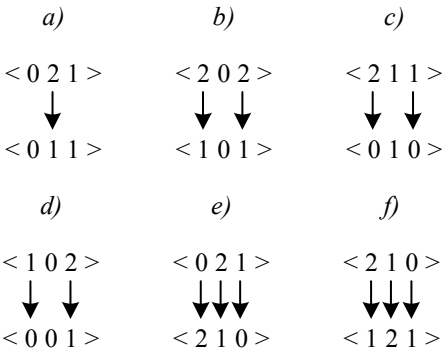
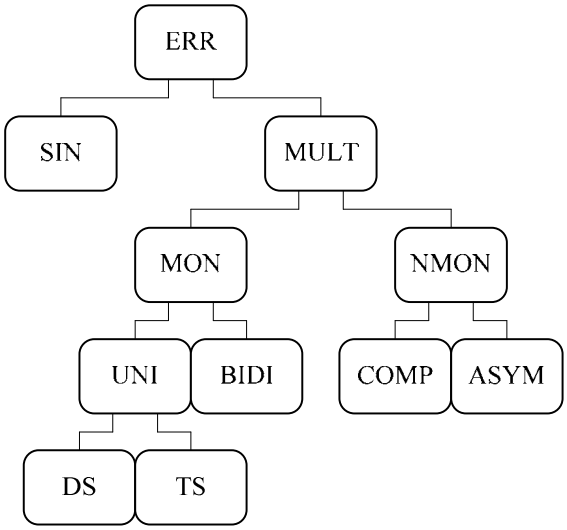


Fig. 4. The examples of various types of errors in ternary code vectors: a) single (SIN); b) unidirectional double-side (UNI DS); c) unidirectional triple-side (UNI TS); d) bidirectional (BIDI); e) compositional (COMP); f) asymmetrical (ASYM).

The Table 3 gives an error characteristic for ternary code vectors with the length *m=3*, where the number of errors of various types and different multiplicities is also indicated, and the indicator τ is calculated. A significant share of errors (54.701%) is occupied by various monotonous errors, while compositional errors account for 9.402%, and asymmetrical errors account for 12.82%. The distribution of errors by types for ternary code vectors differs significantly from the distribution of errors by types for binary code vectors. As the *m* value increases, the error distributions by type and multiplicity change.

TABLE III. THE DISTRIBUTION OF ERRORS BY TYPE AND MULTIPLICITY IN TERNARY CODE VECTORS WITH THE LENGTH M=3

Type of error	Error multiplicity			Total	τ, %
	1	2	3		
SIN	162	–	–	162	23.077
UNI DS	–	51	24	75	10.684
UNI TS	–	54	0	54	7.692
BIDI	–	165	90	255	36.325
COMP	–	36	30	66	9.402
ASYM	–	–	90	90	12.82
Total	162	306	234	702	100



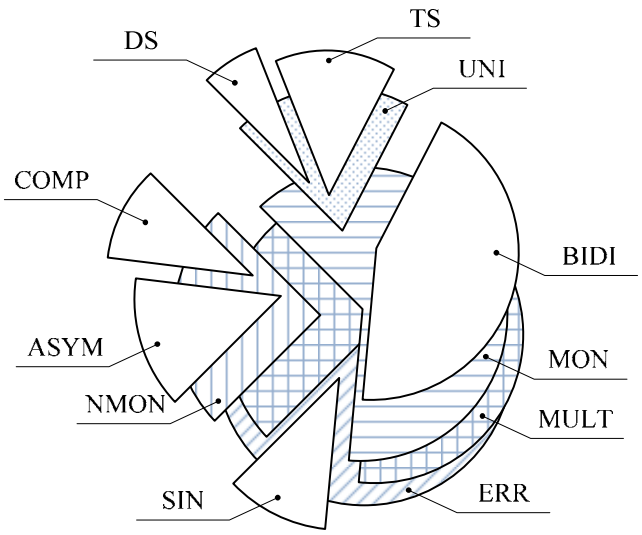


Fig. 3. The classification of errors in ternary vectors.

III. The Codes with Certain Types Error Detection

The redundant codes focused on detecting specific types of errors in code vectors can be used to solve the problems of the synthesis of automation devices with fault detection in both binary and ternary techniques.

The Fig. 5 presents a classification of binary redundancy codes oriented to error detection and used in the construction of discrete systems. The special classes of any unidirectional asymmetrical error detection codes (*UAED*-codes) and any unidirectional error detection codes (*UED*-codes) are stand out among the set of binary redundant codes. In addition, it is necessary to note such codes that detect any unidirectional errors up to a certain multiplicity of d_υ (d_υ-*UED*-codes), and codes that detect any unidirectional and asymmetric errors up to certain multiplicities of d_υ and d_α (d_υ,d_α-*UAED*-codes). There are also narrower classes of codes with the any symmetrical error detection (*SED*-codes) and codes with the any symmetrical error detection up to a certain multiplicity of d_σ (d_σ-*SED*-codes), and codes with the any asymmetrical error detection (*AED*-codes), and codes with the any asymmetrical error detection up to a certain multiplicity of d_σ (d_α-*AED*-codes).

The above classification is due to the implementation features of automation devices with the checkable structures. For example, *UAED*-codes are effectively used for monitoring of automation devices, the outputs of which form monotonously and asymmetrically independent groups (MAI- groups). The *UED*-codes are effectively used for monitoring of automation devices, the outputs of which form a monotonously independent group (MI-groups). There are methods for converting the structures of automation devices into devices with structures whose outputs form MAI- and MI- groups [18]. The codes of the d_υ-*UED* and d_υ,d_α-*UAED* type are used for monitoring automation devices in several MAI- and MI- groups.

The ternary redundant codes can also be classified in the same way. In this case, analogues of binary redundant codes with certain properties can be constructed (Fig. 6). Thus, in the set of ternary redundant codes, we can note codes with any monotonous and asymmetrical error detection. Let's denote them as *MAED*-codes (the difference in notation is due to the broader concept of "monotonicity" in ternary logic and the allocation of both unidirectional and bidirectional monotonous errors). In addition to *MAED*-codes, it is possible to note the codes with any monotonous error detection (*MED*-codes). Such codes can be effectively used in the control of ternary logic devices, the outputs of which form MI- and MAI-groups. To control devices for several groups of monotonously and asymmetrically independent outputs, codes with the monotonous and asymmetrical error detection up to certain multiplicities d_μ and d_α (d_μ,d_α-*MAED*-codes / d_μ-*MED*-codes) can be used. It is also possible to note the ternary codes with any compositional error detection (*CED*-codes) and the codes with any compositional error detection up to a certain multiplicity of d_κ (d_κ-*CED*-codes), and also the codes with any asymmetrical error detection (*AED*-codes) and the codes with any asymmetrical error detection up to a certain multiplicity of d_α (d_α-*AED*-codes). In addition, it should be noted that it is possible to construct the codes focused on detecting only one of the subclasses of monotonous errors: *UED*-codes (d_υ-*UED*-codes) – the codes with any unidirectional monotonous error detection (unidirectional monotonous error detection up to a certain multiplicity of d_υ), *BED*-codes (d_β-*BED*-codes) – the codes with any bidirectional monotonous error detection (bidirectional monotonous error detection up to a certain multiplicity d_β).

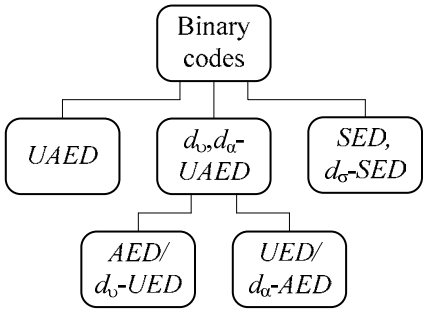


Fig. 5. The classification of binary redundancy codes.

The ternary codes that have the properties of certain type's error detection can be constructed by analogy with binary redundant codes.

A *MAED*-code is a ternary code composed of code words having the same number of bits equal to 1 and 2. This code is called *a constant-composition code* [27, 28]. An ana-

log of this code in the binary logic is the constant-weight code [29]. Here is an example of a compositional code ($C(r_1,r_2)$-code, where r_1 and r_2 are the number of bits equal to 1 and the number of bits equal to 2) with the number of bits m=4. For example, C(1,2)-code form the code vectors from the set {0122, 1022, 0212, 1202, 0221, 1220, 2012, 2102, 2021, 2120, 2201, 2210}.

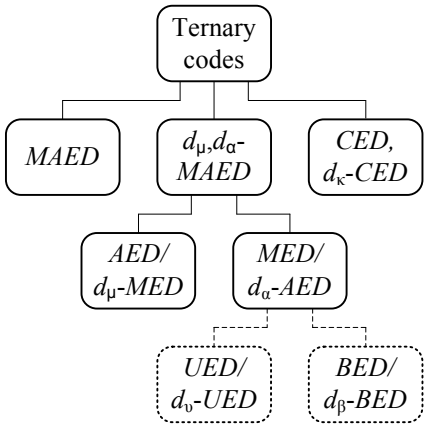


Fig. 6. The classification of ternary redundant codes.

The example of a *MAED*-code is a ternary sum code, which is constructed according to the following rules. In the data vector, we determine the number of bits equal to 1 and the number of bits equal to 2 (the numbers r_1 and r_2 are calculated). The numbers r_1 and r_2 are represented in ternary form and are written respectively in k_1 high and k_2 low order bits of check vectors with the length $k = k_1 + k_2$. The number of bits equal to 1 and the number of bits equal to 2 in the data vector with the length m can be equal from 0 to m. It follows that $k_1 = k_2 = \lceil \log_3(m+1) \rceil$ and $k = 2\lceil \log_3(m+1) \rceil$. The Table 4 provides the calculations of the total number of check bits in ternary sum codes with the different lengths of the data vector. Here is an example of determining the values of the check vector bits for the data vector <011120222> of the ternary sum code: r_1=3, r_2=4, $[r_1]_3$=010, $[r_2]_3$=011. The check vector will have the following bits <010011>.

The sum ternary code, constructed according to the above rules, will have the same check vector for each data vector with the same composition 1 and 2. Thus, the ternary sum code will be analogous to the classical binary Berger code in its features [30].

Note that all *MAED*-codes will have only compositional errors in the class of undetectable errors, while other types of errors will be detected by them. If this property indicates a significant share of undetectable errors of their total number in binary logic [31], then their share will be smaller in ternary logic (because there are significantly more monotonous errors in ternary vectors than in binary ones).

d_μ-*MED*-codes are sum codes, for which the numbers r_1 and r_2 are calculated in the $M \in \{3^1, 3^2, ..., 3^{\lceil \log_3(m+1) \rceil - 1}\}$ modulus residue ring (these are analogs of sum binary modular codes [32]). One of these codes is the code for which the numbers r_1 and r_2 are calculated in the M=9 modulus residue ring. The number of check bits in such ternary modular codes will always be equal to k=4. We obtain the check vector for the data vector <011120222> of the ternary sum code in the M=9 modulus residue ring: r_1(mod9)=3, r_2(mod9)=4, $[r_1]_3$=10, $[r_2]_3$=11. The check vector has the following bits <1011>. The considered code will detect any monotonous error in the data vectors of the multiplicity d_μ<9. This explains the prospects for applying of these codes.

TABLE IV. THE NUMBER OF CHECK BITS IN THE TERNARY SUM CODES

m	k
4	4
5	4
6	4
7	4
8	4
9	6
10	6
…	…
20	6
…	…
50	8
…	…
100	10

Another d_μ-*MED*-codes are codes obtained by combining the codewords of codes with a constant composition of values:

$$C(r_1, r_2, d_\mu) = \bigcup_{r_1, r_2 = \lfloor m/2 \rfloor (\mathrm{mod}\, d_\mu)} C(r_1, r_2), \qquad r_1 + r_2 \le m. \quad (7)$$

For example, let m=8 and d_μ=4, then $r_1, r_2 = \lfloor 8/2 \rfloor (\mathrm{mod}\, 4) = 4(\mathrm{mod}\, 4) = 0$. Thus, to construct a d_μ-*MED*-code with the indicated parameters, it is necessary to take all code vectors belonging to $C(r_1,r_2)$- codes, for which $r_1, r_2 \in \{0,1,2\}$. Moreover, $r_1 + r_2 \le 8$. The indicated values correspond to codes with a constant composition of values C(0,0), C(0,4), C(0,8), C(4,0), C(8,0), C(4,4).

The ternary code constructed according to the rules presented above will be analogous to the Borden binary code [33].

IV. CONCLUSION

The classifications of errors in binary and ternary code vectors presented in the article make it possible to identify the main classes of binary and ternary redundant codes oriented to detect errors of certain types and multiplicities.

In turn, this makes it possible to reasonably choose one or another coding method in the self-checking digital devices developing and in choosing the methods for implementing their diagnostic support.

Compared to errors in binary code vectors, errors in ternary code vectors are much more diverse. The primary type of errors are monotonous errors, which are also divided into several classes according to the type of signal distortion (unidirectional and bidirectional monotonous errors).

The article also provides the examples of ternary codes related to the introduced classes of *MAED*- и d_μ-*MED*-

codes. These codes are analogues of binary sum codes (Berger codes and modular sum codes).

The use of these codes is perspectival in the synthesis of checkable digital devices of ternary logic, as well as technical means of their diagnostics.

REFERENCES

[1] 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.

[2] J. Connely "Ternary Computing Testbed 3-Trit Computer Architecture", California Polytechnic State University of San Luis Obispo, August 29th, 2008, 184 p.

[3] 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.

[4] 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.

[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] 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.

[8] 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/ASPDAC.2018.8297369.

[9] 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.

[10] 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.

[11] D.V. Efanov "Ternary Parity Codes: Features", Proceedings of 17th IEEE East-West Design & Test Symposium (EWDTS`2019), Batumi, Georgia, September 13-16, 2019, pp. 315-319, doi: 10.1109/EWDTS.2019.8884414.

[12] 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.

[13] 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.

[14] 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.

[15] E.S. Sogomonyan, and E.V. Slabakov "Self-Checking Devices and Fault-Tolerant Systems" (in Russ.), Moscow: Radio & Communication, 1989, 208 p.

[16] E. Fujiwara "Code Design for Dependable Systems: Theory and Practical Applications", John Wiley & Sons, 2006, 720 p.

[17] S.J. Piestrak "Design of Self-Testing Checkers for Unidirectional Error Detecting Codes", Wrocław: Ofiyna Wydawnicza Politechniki Wrocławskiej, 1995, 111 p.

[18] 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.

[19] D.V. Efanov, V.V. Sapozhnikov, and Vl.V. Sapozhnikov "Sum Codes with Fixed Values of Multiplicities for Detectable Unidirectional and Asymmetrical Errors for Technical Diagnostics of Discrete Systems", Automation and Remote Control, 2019, Vol. 80, issue 6, pp. 1082-1097, doi: 10.1134/S0005117919060079.

[20] D.A. Pospelov "Logical Methods of Analysis and Synthesis of Circuits" (in Russ), Moscow: Energy, 1974, 368 p.

[21] S. Ghosh, S. Basu, N.A. Touba "Synthesis of Low Power CED Circuits Based on Parity Codes", Proceedings of 23rd IEEE VLSI Test Symposium (VTS'05), 2005, pp. 315-320.

[22] P. Kubalík, and H. Kubátová "Parity Codes Used for On-Line Testing in FPGA", Acta Polytechnika, 2005, Vol. 45, No. 6, pp. 53-59.

[23] R. Ubar, J. Raik, 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.

[24] A. Morosow, V.V. Sapozhnikov, Vl.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.

[25] Vl.V. Saposhnikov, A. Dmitriev, M. Goessel, and V.V. Saposhnikov "Self-Dual Parity Checking – a New Method for on Line Testing", Proceedings of 14th IEEE VLSI Test Symposium, USA, Princeton, 1996, pp. 162-168.

[26] V.V. Sapozhnikov, Vl.V. Sapozhnikov, and D.V. Efanov "Hamming Codes in Concurrent Error Detection Systems of Logic Devices" (in Russ.), St. Petersburg: Nauka, 2018, 151 p.

[27] M. Svanström "A Lower Bound for Ternary Constant Weight Codes", IEEE Transactions on Information Theory, 1997, vol. 43, pp. 1630-1632.

[28] 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.

[29] C.V. Freiman "Optimal Error Detection Codes for Completely Asymmetric Binary Channels", Information and Control, 1962, Vol. 5, issue 1, pp. 64-71, doi: 10.1016/S0019-9958(62)90223-1.

[30] J.M. Berger "A Note on Error Detecting Codes for Asymmetric Channels", Information and Control, 1961, vol. 4, issue 1, pp. 68-73, doi: 10.1016/S0019-9958(61)80037-5.

[31] V.V. Sapozhnikov, Vl.V. Sapozhnikov, and D.V. Efanov "Codes with Summation Detecting Any Symmetric Errors" (in Russ.), Electronic Modeling, 2017, Vol. 39, issue 3, pp. 47-60.

[32] 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.

[33] J.M. Borden "Optimal Asymmetric Error Detecting Codes", Information and Control, 1982, Vol. 53, Issue 1-2, pp. 66-73, doi: 10.1016/S0019-9958(82)91125-1.

Similarity–Difference Analysis and Matrix Fault Diagnosis of SoC-components

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

Mikhail Karavay
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
Moscow, Russia
mkaravay@yandex.ru

Vladislav Sergienko
Design Automation Department
Kharkov National University of Radio Electronics
Kharkov, Ukraine
serhiienko.w@gmail.com

Svetlana Chumachenko
Design Automation Department
Kharkov National University of Radio Electronics
Kharkov, Ukraine
svetachumachenko@icloud.com

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

Hanna Khakhanova
Design Automation Department
Kharkov National University of Radio Electronics
Kharkov, Ukraine
ann.hahanova@gmail.com

Tariq Hama Salih
Design Automation Department
Kharkov National University of Radio Electronics
Kharkov, Ukraine
tareq.it@gmail.com

***Abstract*— Universal metric of data search in cyberspace is proposed; it is based on the use of similarity – difference parameters and matrix structure in binary form. Method for analyzing matrix data structures using the similarity–difference metric for fault detection in digital systems is described. The difference diagnostic method is characterized by the execution of three logical operations on the binary states of the matrix vector-rows and is focused on single and multiple faults in digital systems and software applications. The qubit-difference method for fault detection is characterized by the use of vector parallel logical operations to form a diagnosis of a multivalued technical state, according to the principle of "divide and unite, eliminating contradictions." The advantage of the qubit-vector representation of data in matrix cells is shown, which makes it possible to perform logical operations in parallel in three automatic cycles in order to obtain the required result. Method for detecting single and multiple faults in digital software and hardware systems is described, focused on the hardware implementation of parallel logical register operations, which provide a significant increase in performance compared to existing analogues. Method and hardware implementation of a sequencer for determining similarity–difference–inclusion is proposed, which is characterized by obtaining a more accurate structured assessment of the interaction of two objects.**

Keywords— single and multiple faults, fault diagnosis, similarity–difference metric, qubit vectors, matrix data structures, digital systems-on-chips

I. State of the Art

The hardware and software of a computer at each stage of its development constitute a harmonious alliance. This means that new software systems must correspond to new chips, as well as improved silicon chips and structures must be developed for new algorithms and technologies. It's no secret that the world's leading company Apple uses a strategy of dying old hardware by forcibly not supporting it with new software systems. The situation is similar to new hardware and old applications. The growing computational power of new chips provides scientists with the opportunity to design improved models, methods and algorithms for solving current market and scientific problems. Every programming specialist who wants to be in demand should have an idea of the breakthroughs of a new generation of chips, computers, networks, data centers and cloud services. The best solutions in the field of computing, analysis and data retrieval indicate the existence of a stable trend towards the intellectualization of hardware by implementing hardware solutions in artificial intelligence and machine learning algorithms, which create tools for high-performance computing required today in the cyber-physical space for analyzing large data. It is well known that all models, methods, algorithms and applications related to artificial intelligence tend in their development from the probabilistic characteristics of processes to the determinism of finite state machines originally incorporated in modern computing. It is known that the ideal description of any process or phenomenon has always been, is and will be a strictly deterministic truth table of n variables, as a certain limit of knowledge about the object of research. Therefore, the ways of solving urgent problems of recognizing patterns, images, states will be formed in opposite directions: 1) From probabilistic ignorance of the process to its automaton determinism through the training time. 2) From computer determinism, as the core of some knowledge, to multi-parameter detailing of the analyzed process or phenomenon. The second way involves the development of an effective similarity-difference metric, data structures and algorithms for the analysis and training of initially deterministic systems [1-11]. The goal is to significantly reduce the computational complexity of data retrieval and fault diagnosis algorithms by developing an efficient matrix data infrastructure for technological hardware-focused parallel fault analysis using the similarity-difference metric. Objectives are the following: 1) Development of theoretical foundations of data retrieval based on metric

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

estimation of the interaction of vectors encoding technical (diagnostic) states of a process or phenomenon. 2) Development of a matrix model for describing the infrastructure of processes or phenomena identification (diagnosis). 3) Development of efficient computational algorithms (difference, qubit and equivalence) for data (fault) detection by evaluating the similarity–difference between the rows–columns of the matrix. 4) Coding algorithms and testing them based on the given matrices for diagnosing digital systems. The essence of the research is the creation of computing for accurate data retrieval based on the analysis of regular matrix structures used to create methods and algorithms for parallel partitioning and combining of row and column vectors according to the similarity–difference metric.

II. Unity of Similarity–Difference of Objects or Components

The metric of similarity–difference between objects, processes or phenomena [7], represented by the equation S⊕D=a∪b=U=1, is the basis for solving the most common problems in the technological market related to measurements in cyber-physical space and the Internet. Such scientific and practical problems (Fig. 1) are the following ones: 1) Recognition of images and patterns. 2) Machine learning. 3) Decision-making. 4) Management of processes and objects. 5) Big data analytics. 6) Digitalization of processes and phenomena. 7) Testing and diagnosis of systems [8-10]. 8) Creation and use of regular databases.

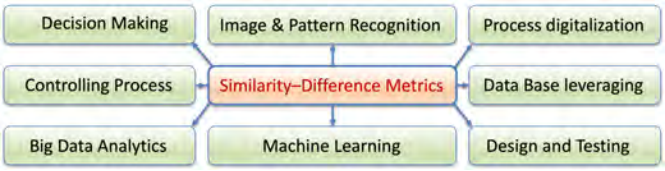


Fig. 1. Scientific and practical problems which can be solved using the similarity-difference metric

The model of relations between processes and/or phenomena is optimally represented in the form of a binary matrix M=[M_{ij}], which in the general case forms a Cartesian product of two sets, for example, a set of tests by a set of functionalities: <T×F>. Similarity is the ratio of the number of identical components of a process or phenomenon to their total number in the metric of the specified essential parameters. The difference is the ratio of the number of different components of a process or phenomenon to their total number in the metric of specified essential parameters. Similarity and difference are mutually complementary assessments of the relationship between processes and phenomena. To determine a scalar estimate of the similarity between all pairs of components, it is necessary to generate a quadratic matrix of their similarity (T×T or F×F) from the matrix M by using a formula that determines the similarity ratio of the two components:

$$M_{ik} = S(M_i, M_k) = \frac{\Sigma_{j=1}^{n}(M_{ij} \wedge M_{kj})}{\Sigma_{j=1}^{n}(M_{ij} \vee M_{kj})};$$

$$D(M_i, M_k) = 1 - S(M_i, M_k).$$

III. Cubit-difference fault detection method

A qubit is defined here as a binary vector of dimension k, simultaneously identifying a finite number of states, as a superposition of their unitary codes. A logical function qubit is a vector that forms k=2^n output states, where n is the number of variables [11,12]. A fault qubit is a vector that forms k faulty state codes of an individual line, component or system that is characterized by the ability to superimpose due to the unitarity of their codes. Test results parameter R=($R_1, R_2, \dots, R_i, \dots, R_p$) defined in the alphabet {0,1} and equal to the length of the test p (where T=($T_1, T_2, \dots, T_i, \dots, T_p$)) is a vector-qubit of the values of the fault function, which is formed in the process of performing a diagnostic experiment. Therefore, any row of the fault matrix M=($M_1, M_2, \dots, M_i, \dots, M_p$) functionally dependent on the vector R: $M_i = f(R = \{0,1\})$ in the process of matrix analysis. It should be noted that the matrix row M_i and the test row T_i hereinafter are equivalent notions. A qubit-difference method for detecting multiple faults D_m is proposed; it is based on calculating the set-theoretic difference of two matrix row vectors corresponding to the union of unit and zero responses of the observed outputs on the input fault test:

$$D_m = \bigcup_{\forall R_i=1} M_i \setminus \bigcup_{\forall R_i=0} M_i = \bigvee_{\forall R_i=1} M_i \wedge \overline{\bigvee_{\forall R_i=0} M_i}.$$

The alphabet for describing stuck-at faults of a digital circuit [11–16] has the following form: A= {0,1,X={0,1},∅}, where the symbol codes (qubits) constitute the set: K(A)={10,01,11,00}. Data structures are represented by a matrix of faults on the Cartesian product of a set of test patterns and a set of equipotential lines of the object under diagnosis, where each cell is a two-bit code – qubit: the first bit identifies the stuck-at zero being detected, and the second one identifies the stuck-at one. Superposition of faults (two 1s on one cell-line) makes it possible to significantly minimize data structures for storing information in order to subsequently detecting faults when performing a diagnostic experiment online. To test the qubit method for fault detection, a logical circuit is proposed below in Fig. 2, which has 6 and-not elements, 11 lines, 5 inputs and two outputs.

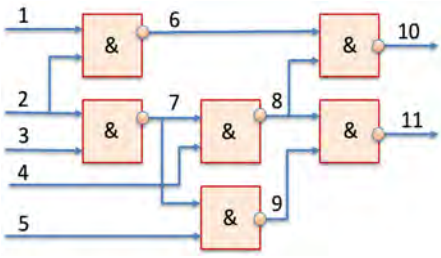


Fig. 2. Logical circuit for the verification of the qubit method

The synthesis of a test for a digital circuit is performed by activating one-dimensional paths, the number of which in this example is five. The following table contains a complete test for detecting single stuck-at faults, which are the negated to fault-free states of activation lines, where empty coordinates correspond to the empty set. In addition, the result of the diagnostic experiment for the output response vector R= 10100100 is shown:

M=<T,F>	1	2	3	4	5	6	7	8	9	10	11	R
T1=	1					0				1	1	1
T2=	0					1				0	1	0
T3=		1					0		1	1	0	1
T4=		0					1		0	1	1	0
T5=			1				0		1	0	0	0
T6=			0				1		0	0	1	1
T7=				1				0		1	0	0
T8=				0				1		0	0	0
T9					1				0	0	1	1
T10					0				1	0	0	0
$F_m^1=\bigvee_{\forall R_i=1} T_i$	1	1	0		1	0	X		X	X	X	
$F^0=\bigvee_{\forall R_i=0} T_i$	0	0	1	X	0	1	X	X	X	X	X	
$D_m=F^1 \wedge \overline{F^0}$	1	1	0		1	0						

Further, we propose a table obtained by unitary coding of fault symbols [11]. It contains a complete test for detecting single stuck-at faults. The table of faults shows technological matrix data structures, and also the execution of a diagnostic experiment based on uniting a set of faults-rows and qubits in cells, which form incorrect states of outputs on test patterns {T1-R10; T5-R11; T6-(R10, R11); T8-R11}. The result of performing a diagnostic experiment for the output response vector R= 10100100 is also shown:

M=<T,F>	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	R10	R11
T1	01	10	01	00	10	00	10	00	10	10	01	1	0
T2	10	00	10	00	01	10	00	00	10	01	10	0	0
T3	00	01	01	00	00	01	10	01	01	10	10	0	0
T4	10	00	01	00	10	00	01	00	10	01	01	0	0
T5	00	10	00	01	00	01	00	10	00	10	01	0	1
T6	01	10	00	00	10	00	00	01	10	01	10	1	1
T7	01	00	00	10	00	00	01	00	10	01	01	0	0
T8	00	10	10	01	01	10	00	00	00	01	10	0	1
$F_m^1=\bigvee_{\forall R_i=1} T_i$	01	10	11	01	11	11	10	11	10	11	11	1	1
$F^0=\bigvee_{\forall R_i=0} T_i$	11	01	11	10	11	11	11	01	11	11	11	0	0
$D_m=F^1 \wedge \overline{F^0}$	00	10	00	00	10	00	00	00	10	00	00		
$D_m=$	.	0	.	1	.	.	.	0	.	.	.		

Here, parallel execution of the disjunction operation for rows T1, T5, T6, T8 forms a vector F_m^1, which collects all possible faults detected on test patterns. A vector F^0 obtained by parallel disjunction operation over rows T2, T3, T4, T7 combines all the impossible, undetectable faults on test patterns. Subtracting all impossible faults (undetectable on the test vectors) from all possible ones gives the desired result in the form of three stuck-at-faults coded in the table as F2 =10; F4=01; F8=10. Thus, the parallel execution of two register Or-operations based on the results of the diagnostic experiment made it possible to determine three possible faults, each of which can be in the logical circuit: $D_m = \{2^0, 4^1, 8^0\}$. A more stringent condition is the existence of a single stuck-at-fault in the logic circuit, which is more likely during the operation of a digital product. The use of such a condition leads to a fault detection procedure based on the following expression:

$$D_s = \bigcap_{\forall R_i=1} M_i \setminus \bigcup_{\forall R_i=0} M_i = \bigwedge_{\forall R_i=1} M_i \wedge \overline{\bigvee_{\forall R_i=0} M_i}.$$

The application of this formula significantly clarifies the diagnostic result in the direction of decreasing the power of possible faults: $D_s = \{2^0\}$ through obtaining a logical contradiction of fault codes based on the use of the And-operation in the columns 4 and 8 in the table below:

M=<T,F>	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	R10	R11
T1	01	10	01	00	10	00	10	00	10	10	01	1	0
T2	10	00	10	00	01	10	00	00	10	01	10	0	0
T3	00	01	01	00	00	01	10	01	01	10	10	0	0
T4	10	00	01	00	10	00	01	00	10	01	01	0	0
T5	00	10	00	01	00	01	00	10	00	10	01	0	1
T6	01	10	00	00	10	00	00	01	10	01	10	1	1
T7	01	00	00	10	00	00	01	00	10	01	01	0	0
T8	00	10	10	01	01	10	00	00	00	01	10	0	1
$F_s^1=\bigwedge_{\forall R_i=1} T_i$	00	10	00	00	00	00	00	00	00	00	00	1	1
$F^0=\bigvee_{\forall R_i=0} T_i$	11	01	11	10	11	11	11	01	11	11	11	0	0
$D_s=F^1 \wedge \overline{F^0}$	00	10	00	00	00	00	00	00	00	00	00		
$D_s=$	.	0	.	.	.	.	.	.	.	.	.		

The intersection of the states-rows according to the conditions of the circuit response to the test determines their similarity, which creates a strict condition for the existence of a single fault in the object.

IV. Sequencer of Hardware Computation of Similarity–Difference–Inclusion of Objects

The goal is to determine the qualitative and quantitative metric interaction between processes and phenomena by using parallel register logical operations. Objectives are the following: 1) Synthesis of a logical structure for calculating the metric interaction between processes and phenomena specified in binary code. 2) Encoding data structures and algorithms using one of the hardware description languages. 3) Analysis and testing of the resulting sequencer structure on a representative sample of test input vectors. The absolute scores of difference and similarity do not make much sense when calculating interactions between sets. More informative are the normalized estimates of similarity–differences, reduced to the denominator in the form of the sum of the essential coordinates of two vectors, excluding only the values of X and Y that are zero in the same coordinates. The denominator for obtaining normalized estimates is obtained by summing all (unit) coordinates after performing the disjunction operation $N=\Sigma_{i=1}^{n}(X_i \vee Y_i)$. The numerator can be Hamming distance, which forms an absolute estimate of the difference between binary vectors. But further included is a modification associated with dividing the similarity or difference by the number of unit coordinates obtained after the logical addition of the binary vectors $X \vee Y$. This addition makes the normalized difference (similarity) score more

significant by decreasing the denominator. Thus, there are two normalized estimates:

$$D^n=\frac{D}{N}=\frac{\Sigma_{i=1}^{n}(X_i\oplus Y_i)}{\Sigma_{i=1}^{n}(X_i\vee Y_i)};\ S^n=\frac{S}{N}=\frac{\Sigma_{i=1}^{n}(X_i\wedge Y_i)}{\Sigma_{i=1}^{n}(X_i\vee Y_i)}.$$

The procedure for calculating the reduced estimates of similarity – difference is reduced to performing three vector parallel operations ($\oplus,\wedge,\vee$) with the subsequent counting of units in the resulting vectors. Qualitative analysis for making a decision to include a key data in the equivalence class is based on estimates of the maximum similarity or minimum difference between two objects-rows. The digital logic diagram for calculating the similarity – difference is shown in Fig. 3. There are 4 layers of data transformation: 1) Formation of input data about objects. 2) Synthesis of a common metric for measuring objects. 3) Unitary coding of objects in the synthesized metric of parameters. 4) Input of vectors corresponding to objects to the SD-automaton. 5) Calculating – forming four outputs-values defining the relationship between objects. The block diagram of a device implemented as SD-automaton (Similarity–Difference) was created in Verilog hardware description language for determining the similarity-difference between two objects.

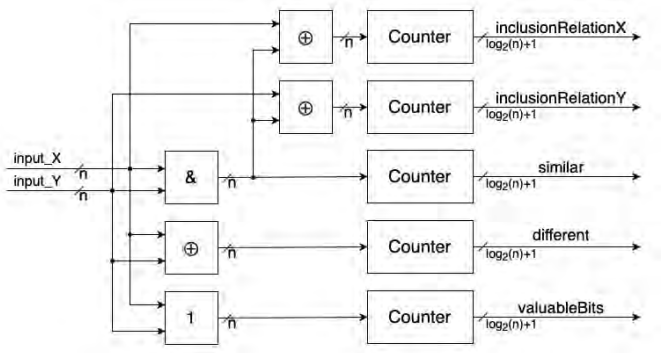


Fig. 3. Block diagram of SD automaton

The device counts the number of matching non-zero bits (similar), different bits (different), significant coordinates (valuableBits), and also the degree of inclusion of X and Y (inclusionRelationX and inclusionRelationY, respectively). Coordinates are considered significant if at least one of the input vectors contains a unit in a bit at the given coordinate. For a more accurate calculation of the similarity/difference, the purely zero coordinates of the same name X and Y are removed – only non-zero ones are taken into account. Counting the number of matching non-zero bits performs in two stages. At the first stage, the "&" operation is performed. After that, the resulting vector of matching non-zero bits is loaded to a counter, where the number of ones in this vector is counted. Depending on the specification of the project, this operation can be performed both directly in this module and on the processor unit, and if not required, it may not be performed at all, since values are often stored in integers. The same goes for the membership metric. The developed module is parameterizable and can be used for any size of input vectors. Below, in the table, there is a listing of verified software module for calculating the similarity – difference of two metrically parameterized objects. Verification of the SD-module software was carried out with several dozen test objects, which were transformed to binary vectors for their subsequent parallel processing on a synthesized digital automaton. As an example, a table is presented below, and also the calculated similarity – difference – inclusion graphs for ten pairs of objects, Fig. 4.

Master, bit	45	89	36	56	77	21	22	18	12	77
Slave, bit	47	72	31	65	78	37	45	11	21	67
Similarity	0,9	0,8	0,6	0,7	1,0	0,5	0,6	0,0	0,8	0,3
Difference	0,1	0,2	0,4	0,3	0,0	0,5	0,4	1,0	0,2	0,7
M⊆S	1	0	0	0	1	0	1	0	0	0
S⊆M	0	0	1	0	1	0	0	0	1	0

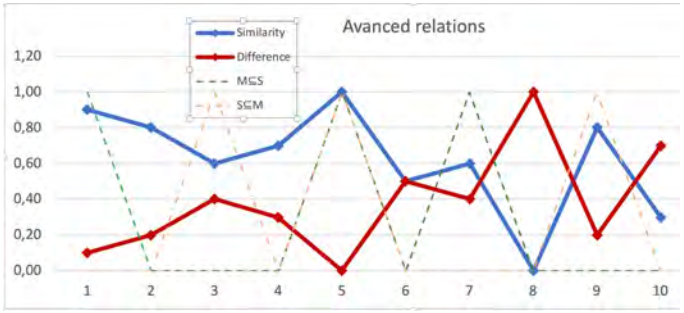


Fig. 4. Verification of similarity–differences–inclusion sequencer

It can be seen from the graph that the symmetry of the similarity – difference relations is not always supported (confirmed) by the graphs of the inclusion relation, which actually make the relations between objects almost always asymmetrical. The top two rows of the table show the number of essential parameters for identifying objects. In this case, their different number is reduced to a common metric by combining particular metrics, after which four output values of the similarity–difference automaton is calculated. The last four rows by their estimates form the structure of the interaction of two vectors-objects. The estimates take into account not only the similarity-difference, but also the inclusion relation, which is not obvious from the values of the similarity of objects. The hardware sequencer for searching for similarity–difference allows identifying the structure of relations between objects. Thus, the scientific novelty of the proposed implementation of the method for defining similarity–difference–inclusion lies in obtaining a structured assessment of the interaction of two objects, which makes it possible to more accurately determine the ways of transforming one object into another, and also to select more significant objects of a pair when making a decision.

V. Conclusion

1) Methods for the analysis of matrix data structures by the similarity – difference metric for detecting faults in digital systems have been developed. Methods for the analysis can be used to search for data in arbitrary matrix structures, including those specified by real numbers.

2) The method for the analysis of matrix data structures is presented, which use a binary vector of matrix differentiation (output response vector), and also algorithm for finding the required data by analyzing rows and/or columns. The diagnostic

method is characterized by the execution of three logical operations on the binary states of the vector rows of the matrix and is focused on detecting single and multiple faults in digital systems and software applications. The second, The qubit-difference method for detecting faults is characterized by the use of only three vector parallel logical operations to form a diagnosis of a multivalued technical state, according to the principle of "divide and unite, eliminating contradictions."

3) The method for determining the similarity-difference in the multivalued metric of object transformation is formalized, aimed at creating a specialized processor for parallel solving problems of synthesis and analysis of new processes and phenomena through the use of logical operations, which makes it possible to create an optimal strategy for transforming one object into another.

4) A method and hardware implementation of a sequencer for defining similarity-difference-inclusion is proposed, which is characterized by obtaining a structured assessment of the interaction of two objects, which makes it possible to more accurately determine the ways of transforming one object into another, and also to select more significant objects of a pair when making a decision.

5) The limitations of the proposed data retrieval methods are related to the need to perform the procedure for synthesizing matrix data structures, which has a quadratic computational complexity. The implementation of data retrieval and diagnostics methods is implemented in the educational process. Further research will be focused on creating a software and hardware product that implements parallel data retrieval procedures in a metric matrix space based on a given vector of input conditions.

REFERENCES

[1] S.K. Moore, "96-Core Processor Made of Chiplets," [https://spectrum.ieee.org/tech-talk/semiconductors/processors/core-processor-chiplets-isscc-news]

[2] S.K. Moore, "Edge-AI Startup Gets $60-million, Preps for Mass Production," [https://spectrum.ieee.org/tech-talk/semiconductors/processors/israeli-edgeai-startup-gets-60million-preps-for-mass-production]

[3] S.K. Moore, "Cerebras's Giant Chip Will Smash Deep Learning's Speed Barrier," [https://spectrum.ieee.org/semiconductors/processors/cerebrass-giant-chip-will-smash-deep-learnings-speed-barrier]

[4] T.S. Perry "How the Father of FinFETs Helped Save Moore's Law," [https://spectrum.ieee.org/semiconductors/devices/how-the-father-of-finfets-helped-save-moores-law]

[5] C.Q. Choi, "Image Sensor Doubles as a Neural Net", [https://spectrum.ieee.org/tech-talk/computing/hardware/image-neural]

[6] M. Abramovici, M.A. Breuer and A.D. Friedman, Digital System Testing and Testable Design. Comp. Sc. Press, 1998.

[7] M. Karavay, V. Hahanov, E. Litvinova, H. Khakhanova and I. Hahanova, "Qubit Fault Detection in SoC Logic," 2019 IEEE East-West Design & Test Symposium (EWDTS), Batumi, Georgia, 2019, pp. 1-7.

[8] J. Drozd, A. Drozd, M. Al-dhabi, "A resource approach to on-line testing of computing circuits," IEEE East-West Design & Test Symposium, Batumi, Georgia, 2015, pp.276-281.

[9] O. Drozd, M. Kuznietsov, O. Martynyuk, M. Drozd, "A method of the hidden faults elimination in FPGA projects for the critical applications," 9th IEEE International Conference DESSERT, Kyiv, Ukraine, 2018, pp. 231-234.

[10] O. Drozd, V. Antoniuk, V. Nikul, M. Drozd, "Hidden faults in FPGA-built digital components of safety-related systems," 14th International Conference "TCSET'2018, Lviv-Slavsko, Ukraine, 2018, pp. 805-809.

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

[12] I. Hahanov, W. Gharibi, I. Iemelianov, T.B. Amer, "QuaSim – Cloud Service for Digital Circuits Simulation," Proceedings of IEEE East-West Design & Test Symposium, 2016, Yerevan, Armenia, pp. 363-370.

[13] V.I. Hahanov, T.B. Amer, S.V. Chumachenko, E.I. Litvinova, "Qubit technology analysis and diagnosis of digital devices," Electronic modeling, 2015, Vol. 37, no 3, pp. 17-40.

[14] V. Hahanov, E. Litvinova, W. Gharibi, S. Chumachenko, "Big Data Driven Cyber Analytic System," IEEE International Congress on Big Data, New York City, 2015, pp. 615-622.

[15] V. Hahanov, M. Bondarenko, E. Litvinova, "The Structure of a Logical Associative Multiprocessor," Automation and Remote Control, 2012, № 10, pp. 73-94 (In Russian).

[16] V. Hahanov, S. Chumachenko, E. Litvinova, V.H. Abdullayev, A. Hahanova, T. Soklakova, "Cyber Social Computing," 2018 IEEE East-West Design & Test Symposium (EWDTS), 2018.

FPGA Implementation of a Low Latency and High SFDR Direct Digital Synthesizer for Resource-Efficient Quantum-Enhanced Communication

N. Fajar R. Annafianto*[,1], I.A. Burenkov[2], H.F. Ugurdag[1], and S.V. Polyakov[2]
[1]Electrical and Electronics Engineering Dept., Ozyegin University, Istanbul, Turkey
[2]University of Maryland, Maryland, USA
*annafianto.annafianto@ozu.edu.tr

***Abstract*—A Direct Digital Synthesizer (DDS) generates a sinusoidal signal, which is a significant component of many communication systems using modulation schemes. A CORDIC algorithm offers minimum memory requirement compared to look-up-based methods and low latency for this module. The latency depends on the number of iterations, which is determined by the number of angles in the rotation set. However, it is necessary to maintain high spectral purity to optimize the overall system performance. To optimize the opportunity of quantum measurement, low latency and high spectral purity sine wave generator is essential. The proposed design's implementation generates output with 64% latency reduction compared to that of the conventional CORDIC design and 72.2 dB SFDR value.**

Keywords—FPGA, CORDIC, DDS, SFDR, pipeline, latency.

I. Introduction

In most modulation schemes for a digital telecommunication system, a fast and efficient sinusoidal signal generator is needed. Here we report on an FPGA implementation of a versatile Coordinate Rotation Digital Computer (CORDIC) based Direct Digital Synthesizer (DDS). Most commercial lightwave communication systems use standard modulation protocols, such as Phase-Shift Keying (PSK) and Frequency-Shift Keying (FSK), whose implementation is supported by specialized dedicated hardware. There is a need for significant improvement in energy and bandwidth efficiency. Therefore, to gain further improvement a new class of communication systems, namely quantum-measurement enhanced optical communication systems are being actively pursued. In those systems, a classical receiver is replaced with a quantum receiver, while the transmitter remains the same. Recognizing that properties of quantum measurement are in general different from that of classical measurement, more complex modulation schemes than PSK and FSK turns out to be more beneficial [1]. Digital synthesis of these signals requires versatile DDS whose development is reported here. By design, a DDS generates signals with a nearly arbitrary combination of phase and frequency modulations. Many other applications such as software-defined radio, wireless satellite transceiver, HDTV transmission, radar communication, etc. can take advantage of this low latency and re-configurable sine wave generation [2]. Hence, with high spectral purity and low latency, the DDS accommodates the energy-efficient and rapid response properties of quantum measurement instruments to optimize the utilization of the offered capability to surpass that of classical measurement and to maximize the modulation capabilities.

Many strategies and techniques have been developed to enhance the hardware area and speed efficiency of CORDIC algorithm implementation. CORDIC was initially introduced by Volder in 1959 to calculate trigonometric functions in digital hardware devices [2]. Later, a modified version of a CORDIC algorithm was proposed by S. Walther with the ability to calculate circular, hyperbolic, and linear rotation systems [4]. The motivation to use this algorithm in digital platforms has gained popularity since then. Refinements on the efficiency of implementation have resulted in reduced latency and hardware area usage.

In the next section, we provide background information on several CORDIC techniques that are adopted in this work. In section III, we explain the implementation of the proposed method. In section IV, we summarize and compare the results. Finally, we conclude with an evaluation and discuss the prospective developments.

II. Background

The demand for higher-throughput communication has always existed and provides the environment for developing new applications. Enhancing the performance of a DDS enables such throughput increases. Many communication systems use a modulation scheme that generates sinusoidal signal output. One popular approach is to use a DDS hardware module that takes the Frequency Tuning Word (FTW) or Frequency Control Word (FCW) as input and passes the amplitude of the sinusoidal signal to the output. Figure 1 shows the general block diagram of a DDS which consists of a phase accumulator, a signal processor, Digital-to-Analog Converter (DAC) and a low pass filter. The phase accumulator defines the frequency of the sinusoidal signal as the increment of the output phase that is dictated by the input FCW, hence the smaller the input the lower the resulting signal's frequency and vice versa. The low pass filter removes aliasing of the signal of the signal processor's output that contains some noise due to the techniques being employed. Here, we focus on the signal processor that takes phase as the input and generates sinusoidal amplitude.

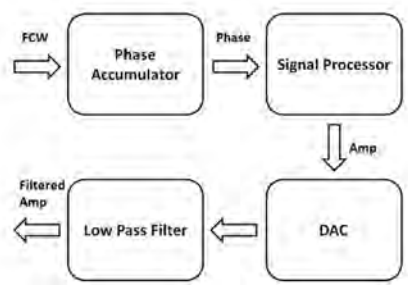


Fig. 1. General block diagram of DDS

CORDIC provides an efficient hardware implementation in terms of area utilization, power consumption, and latency.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

There are three popular approaches to the realization of DDS: Look-Up Tables (LUTs), Polynomial Function, (namely Tylor series expansion), and a CORDIC algorithm [5]. The LUTs occupy memory, namely, Read-Only Memory (ROM), to store the amplitude of the sinusoidal signal. LUTs are computationally fast, but they require a considerable amount of memory even when compression techniques are used [6]. The memory occupancy is mainly based on the width of FCW input and the width of the output. Indeed, to get higher spectral purity, the quantization error is minimized by increasing the LUT memory widths of the output amplitude to yield higher precision. For these reasons, memory size grows significantly with the greater spectral purity requirement. In turn, more memory results in higher power consumption, slower operation, and lower stability [7]. The Taylor series expansion has a complicated implementation that uses several multipliers/dividers. In addition, to get higher spectral purity, higher-order terms should be computed which also means longer latency [5]. A CORDIC algorithm calculates sinusoidal amplitude by a set of rotations. The rotational angles in the set are processed in series and the accumulation of the angles approximates the desired angle that corresponds to the necessary output. The number of angles in the set determines the number of iterations, hence the latency. Thus, the correct selection of an angle set is essential. The rotational operation is carried out by adders, logic shifters, and optionally a small amount of memory that makes implementation and integration easier and simpler [3]. For these reasons, CORDIC advances in resource utilization and power consumption.

Spurious Free Dynamic Range (SFDR) defines the spectral purity of the produced signal. The spectral purity of a signal is significant for the overall performance of the system. Since there exists more than one frequency component in a signal, it is necessary to keep the desired frequency's dominance over the spurious components. SFDR implies the ratio of the power of the desired signal and the power of the strongest spurious signal. Thus, the higher SFDR the smoother the output obtained, which is preferred and pursued in this work.

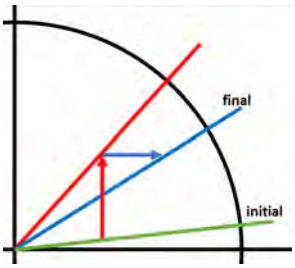


Fig. 2. Rotation mode of CORDIC in DDS with two angular steps

In general, CORDIC has two functional modes: a vector mode and a rotation mode. Our DDS algorithm is based on rotation mode. In rotation mode, the initial vector experiences several rotations in cartesian coordinates based on the angle set and reaches the desired vector position that corresponds to the destination phase. Figure 2 shows the rotation mode with two phases in the set. In the vector mode, the destination angle is estimated by using a set of pre-specified vectors as the reversal of the rotation mode, where it uses the vectors to approximate the target angle [5]. However, CORDIC also comes with some drawbacks. First, it needs scale factor compensation due to numerical operations in the algorithm. Usually, the result is achieved by dividing the scale factor with the output of the series of rotation. To eliminate this requirement, the initial vector is arranged such that it has already been regulated (pre-divided) with the scale factor prior to the calculations. Secondly, accuracy restriction: the number of rotations and the selection of the angles set impacts how close the final angle is to the desired angle [8]. A smaller angle set is beneficial. The following sections describe components in each stage of the hardware architecture and the mathematical operations that they employ.

A. Conventional CORDIC

The first CORDIC algorithm was proposed by removing the burdensome cosine and sine functions multiplications with logic shift operations. Equation (1) computes a rotation of the initial vector (x, y) to the vector (x', y') with the angular distance of θ.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (1)$$

$$\theta = \sum_{i=0}^{b-1} \alpha_i \quad (2)$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \prod_{i=0}^{b-1} cos\alpha_i \begin{bmatrix} 1 & -d_i \tan\alpha_i \\ d_i \tan\alpha_i & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (3)$$

The angle set of α_i converges to θ with a combination of clockwise and/or counterclockwise rotations, see equation (2).

Substituting (2) into (1) and taking $cos\alpha_i$ out of the matrix, we obtain equation (3), where d_i corresponds to the direction of rotation at the respective stage i, given as $d_i = \{-1, 1\} = sign(z_i)$, -1 is for counterclockwise direction and 1 is for clockwise direction. b is the angle set size and the number of iterations. z_i is the remaining phase at iteration i. Our goal is to establish a recurrent formula for rotations that can be conveniently calculated on an FPGA.

$$\alpha_i = \arctan(2^{-i}) \quad (4)$$

$$\prod_{i=0}^{b-1} cos\alpha_i = K \quad (5)$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = K \begin{bmatrix} 1 & -d_i * 2^{-i} \\ d_i * 2^{-i} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (6)$$

$$z_i = \theta - \sum_{c=i}^{b-1} \alpha_c \quad (7)$$

K in the equation (5) is the overall scale factor that can be pre-calculated for the initial vector (x, y). Substituting equation (4) and (5) into equation (3), we obtain equation (6). The division by 2^i can be replaced by an arithmetic shift operator. Thus, iterations are written as:

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} 1 & -d_i * 2^{-i} \\ d_i * 2^{-i} & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad (8)$$

The block diagram in Figure 3 shows three stages (i-1), (i), and (i+1) of a conventional CORDIC as the realization of equation (8). The multiplexers have +/- tags that determine the additions or extractions of variables based on the sign of z_i. Note that intersections of lines show the crossing of paths with no connection between them, this is valid for all diagrams.

Equation (4) implies an angle in the angle set for the iteration i. For the sake of simplicity, by selecting 7 iterations

($i = [0, 6]$), we obtain the following angle set: {45, 26.565, 14.036, 7.125, 3.576, 1.789, 0.895}. Note that the conventional CORDIC has 15 iterations. To ensure that z_i is approximately 0 at the end for any destination angle θ, the number of iterations (b in equation (3)) is specified as 15, hence i ranges from 0 to 14. This number is also the accuracy limit of the digital system which uses variables with a bit width of 16 bits, because shifting more than 15 bits results in 0 in such a variable, that variable has no impact on the algorithm, the operations add redundant latency and produce no modification on the final output. However, the range of convergence, that specifies the absolute value of the angle θ, is 99.882. One uses a domain folding technique with 2 blocks that cover [-90, 90) and [90, 270). This way any θ can be reached by locating an appropriate initial vector.

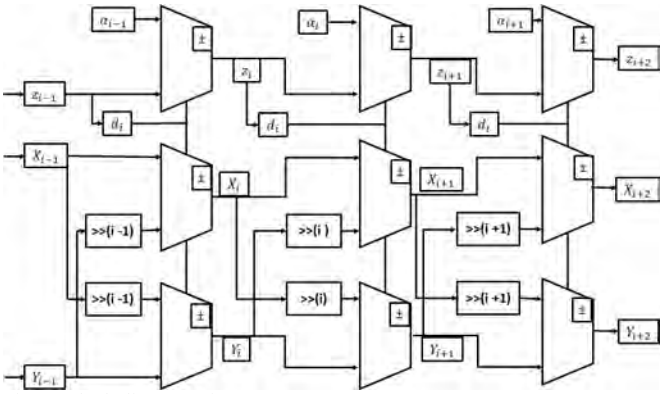


Fig. 3. Block diagram of Conv DDS

B. *Scaling Free CORDIC*

As the name implies, Scaling Free CORDIC pursues an algorithm to avoid multiplication by scale factor prior to the final output for fast performance. Scaling-free CORDIC recognizes one direction of rotation and a halting state, meaning $d_i \in \{0, 1\}$. It rotates counterclockwise only when z_i is greater than the angle at stage i or stays at the current position otherwise. Thus, z_i is always a positive number. This makes the attainable maximum frequency higher as we will see in the result section. The sine and cosine terms can be simplified when any of them is considerably small.

$$\left[\frac{w-log_2 6}{3}\right] \leq j \leq w-1 \tag{9}$$

$$\begin{bmatrix} sin\theta \\ cos\theta \end{bmatrix} = \begin{bmatrix} 2^{-i} \\ 1-2^{-(2i+1)} \end{bmatrix} \tag{10}$$

The approximation in equation (10) is accurate if the requirement in equation (9) is met [4], where w is the bit width. By substituting (10) and (2) into (1) we obtain:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \prod_{i=0}^{b-1} d_i \begin{bmatrix} 1-2^{-(2i+1)} & -2^{-i} \\ 2^{-i} & 1-2^{-(2i+1)} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{11}$$

Then, the recursive formula is given by:

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = d_i \begin{bmatrix} 1-2^{-(2i+1)} & -2^{-i} \\ 2^{-i} & 1-2^{-(2i+1)} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \tag{12}$$

Note that (12) has no scale factor unlike in (6). Figure 4 depicts the block diagram of the Scaling Free CORDIC algorithm at stage i.

The low range of convergence uses a domain folding technique to squeeze the blocks into several extra regions. Due to the condition in (9), and with bit width (w in equation (9)) of 16, where j is i+1, the approximation in (10) only holds for i between 3 to 14, but after 8th iteration, the logic shifter of (2i+1) results in more than 17 bits shift. Thus, iterations 8 through 14 have no effect. Similar to the previous argument, that variable has no further effect on the algorithm, the operations add more latency and produce no modification on the output. Therefore, iterations 9 through 14 are omitted, to reduce latency and redundant area usage. Hence i goes between 3 to 8. The range of convergence becomes [0,22.5). The low range of convergence requires extensive use of a domain folding technique. To obtain convergence, 16 domains folding is employed and requires multiplication by a bothersome factor of $1/\sqrt{2}$. Thus, each domain's distance is 20 degrees which is within the range of convergence.

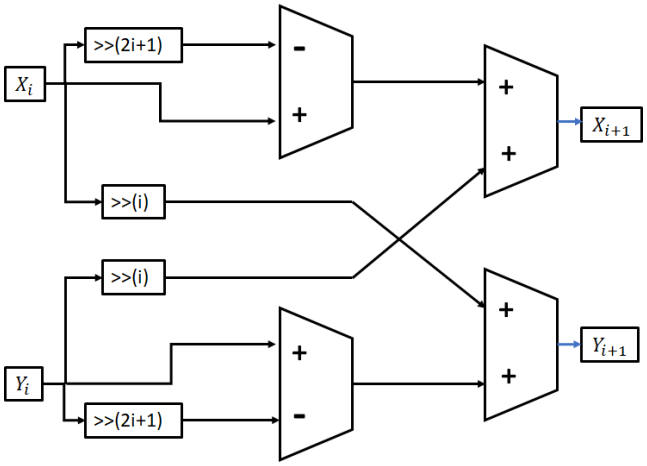


Fig. 4. Block diagram of Scaling-Free CORDIC

The argument reduction technique improves the latency of the method. Particularly, one may jump over several stages by predicting the output at the end of the skipped stages [4]. Typically, more than one computational path is possible, particularly at the early stages. This is because the initial angles are larger than that of the last stages. These computational paths can be pre-computed, and the results can be assigned using multiplexers, probable combinations of output in the earlier stages are still few and can be estimated by using multiplexers. Hence the first 3 stages are skipped and the output at the end of the 3rd stage is obtained. However, the argument reduction technique requires a variadic scale factor, therefore scale factor multiplication is not avoided completely [4]. Nonetheless, this technique reduces a significant amount of latency (from 12 to 9 iterations as we see in the result section for state-machine based design). The consequence of not reaching the desired angle due to the insufficiency of the range of convergence is to repeat iteration for the rest of the angular gap. The angular gap means the remaining angle to the desired angle that the range of convergence couldn't cover. Thus, double and even triple latency may occur. Domain folding and the argument reduction techniques are critical in this regard.

The angle set is {36.87/16.26/0, 7.125/0, 1.789, 0.895, S*0.112} where S is an integer in a range from 0 to 8 [9]. The range of convergence is (-57.57, 57.57). Thus, to cover the entire space, quadrant domain folding is being adopted.

Domain folding occurs at the first stage. The computation of each phase assumes different strategies.

Friend angles: Any group of angles that have identical magnitude is considered friend angles [9]. For instance, in Cartesian coordinate R = 4 + 3i with the phase of 36.869 and R = 5 with the phase of 0 are friend angle because they have the same magnitude of 5. Thus, all angles in Figure 2 are friend angles since they all have the same magnitude of 1. The identical magnitude is essential for the consistency of the system because different magnitudes impose divergence in power gains and result in different scale factors that make the system even more complicated.

Redundant CORDIC: Conventional rotator moves a vector in either direction: clockwise or counterclockwise. However, rotation with a large angular gap may require the next angles to cover up the unnecessarily extensive jump in a reverse direction. In those cases, holding the position instead of rotating is advantageous. Thus, the direction of rotation choices is $d_i = \{-1, 0, 1\}$. However, adding one more "direction", that is 0 or no angular movement, yet regulated with appropriate power gain to attain consistency with the other directions, reduces the maximum frequency of the design.

Nanorotator: Rotation by a sufficiently small angle can be approximated further. Given $R = A + Si$, a rotation is sufficiently small if S<<A, therefore $\alpha = \arctan(S/A) \approx S/A$. The other rotators are the same as previously explained for CORDIC algorithms.

III. Implementation

To ensure a successful implementation, we have chosen the workflow depicted in Figure 5. We implemented the algorithm as a MATLAB script and simulated the code, taking advantage of functions that ultimately are not feasible in the hardware platform, such as floating-point, exponentiation, and numerous other operators. This reduces design effort and completion time. Then, we verify the result and evaluate the performance in the software domain, which gives us an insight into the possible performance in the hardware domain. We write the register transfer level (RTL) implementation of the design on Xilinx ISE using Verilog HDL. Then, we designed the testbench to simulate the RTL implementation and confirm its functionality. Since all variables in the hardware are in integer, we map the hardware simulation results to that of MATLAB simulation for consistency. The hardware platform we utilized is the Xilinx Virtex-6 ML605 FPGA. As a next step, we verify the FPGA's functionality using an Integrated Logic Analyzer (ILA). We store and extract the output values from the ILA signal analyzer, compare the results with that of RTL simulation, and assess the data for evaluation as shown in Figure 10.

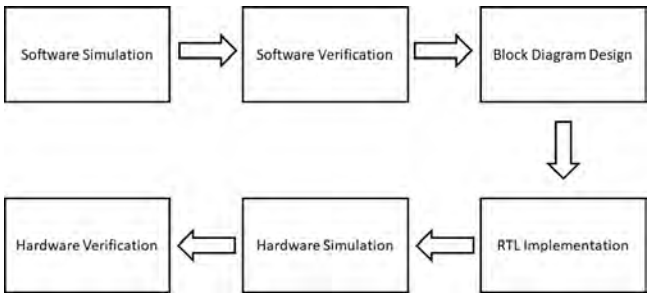


Fig. 5. Workflow

Here, we describe the stages of our implementation:

Stage 1: The domains folding technique, using 8 domains, retains resource efficiency for the system. Additionally, a higher maximum frequency is achieved using one-directional rotations. Folding the coordinate space into several domains leads to a smaller convergence range, but when the number of domains reaches or exceeds 16, complicated operations such as multiplication by $1/\sqrt{2}$ are required. Thus, we use 8 domains to ensure simplicity. The assignment of the initial vector can be done by trivial swapping between imaginary and real parts and negation as we see in Table 1. The angular range of each domain is 45 which is within the convergence range of the angle set in the counterclockwise direction: it enables one-directional rotation for the next stage.

Table 1. Eight domain folding coordinate assignment

Domain	X	Y
0-45	X	Y
45-90	Y	X
90-135	-Y	X
135-180	-X	Y
180-225	-X	-Y
225-270	-Y	-X
270-315	Y	-X
315-360	X	-Y

Stage 2: The first rotation yields 3 phase options with angles {36.87, 16.26, 0}. All rotation coefficients have the same magnitudes that imply the same power gain/scale factor. We use coefficients, with an angular magnitude of 1.5625. Hence, R = 1.25 + 0.9375i for phase of 36.869 degrees, R = 1.5 + 0.4375i for phase of 16.26 degrees, and R = 1.5625 for phase of 0 degrees. Equation (12) is modified to optimize the hardware implementation for the above three angles such as:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 + 2^{-2} & -1 + 2^{-4} \\ 1 - 2^{-4} & 1 + 2^{-2} \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \quad (13)$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2^{-1} + 1 & -2^{-1} + 2^{-4} \\ 2^{-1} - 2^{-4} & 2^{-1} + 1 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \quad (14)$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2^{-1} + 1 + 2^{-4} \\ 2^{-1} + 1 + 2^{-4} \end{bmatrix}^T \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \quad (15)$$

Equations (13), (14), and (15) can be implemented with just logic shifter and adder.

Resource sharing eliminates redundancy in resource usage. In Figure 6, we use 6 logic shifters as some operators share the same logic shifter's output. The switching rules for the multiplexers are shown as numbers {0, 1, 2}, where {0, 1, 2} encodes the jump angle {36.87, 16.26, 0}, respectively. The architecture of stage 2 is somewhat complex, which may impact the maximum frequency of the hardware implementation. Thus, having a one-directional rotation approach shortens the longest path of the architecture, and in our case, we have 3 rotational options instead of 5 in the regular mode, which shrinks the area usage and improves the speed of this segment.

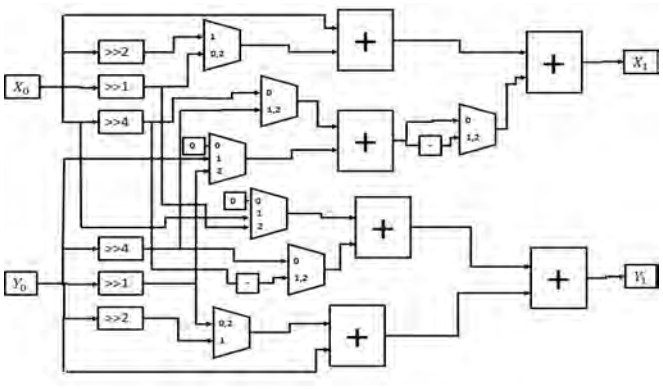


Fig. 6. Stage 2 block diagram

Stage 3: In this stage, we adopt redundant CORDIC to eliminate several rotations and guarantee convergence provided by remaining angles in the set. The coefficients of this rotator have an angular magnitude of 1.0078125: R = 1 + 0.125i for a phase of 7.125 degrees, and R = 1.0078125 for a phase of 0 degrees. Equation (12) turns into:

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & -d * 2^{-3} \\ d * 2^{-3} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \quad (16)$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 + 2^{-7} \\ 1 + 2^{-7} \end{bmatrix}^T \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \quad (17)$$

Hardware implementation of equations (16) and (17) requires just 2 logic shifters per coordinate. The direction d in (16) can be {-1, 1}. The rotation by 0 degrees (described by equation (17)) is equivalent to a no-rotation choice. Such redundancy is tolerable because we end up with three jumping options similar to that of the previous stage. No degradation in the maximum frequency of the design results from this architecture in that regard. The coefficients in stages 2 and 3 ensure consistency of scale factor as the friend angle's condition is fulfilled.

The block diagram for this stage in Figure 7 shows 4 multiplexers where two of them have the tag numbers. 0 indicates the halting condition for no rotation of the current vector. Note that the appropriate power gain is imposed. 1 indicates either clockwise or counterclockwise rotation set by the sign of active phase z.

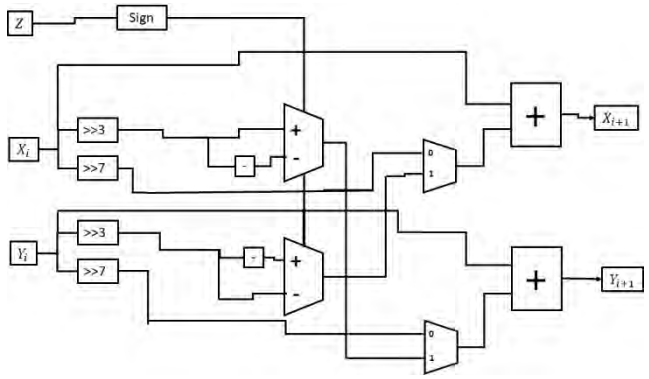


Fig. 7. Stage 3 block diagram

Stage 4: Entering this stage, the residual angle gap's range is 3.58, which is within the range of convergence of the remaining angle in the set. Hence, this stage requires no redundant CORDIC rotation: $d = \{-1, 1\}$. In this stage, we adopt conventional CORDIC architecture at the 5th iteration. The coefficient is R = 1 + 0.03125i for a phase of 1.789 degrees, and the hardware compatible computation is given by equation (18):

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} 1 & -d * 2^{-j} \\ d * 2^{-j} & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad (18)$$

The hardware implementation requires two shifters, Figure 8. For this stage i = 3 and j = 5.

Stage 5: We reuse conventional CORDIC architecture similar to the previous stage but with R = 1 + 0.015625 for a phase of 0.895 degrees. The hardware compatible computation is also given by equation (18), but here i = 4 and j = 6. The block diagram is identical to that of stage 4, which is Figure 8.

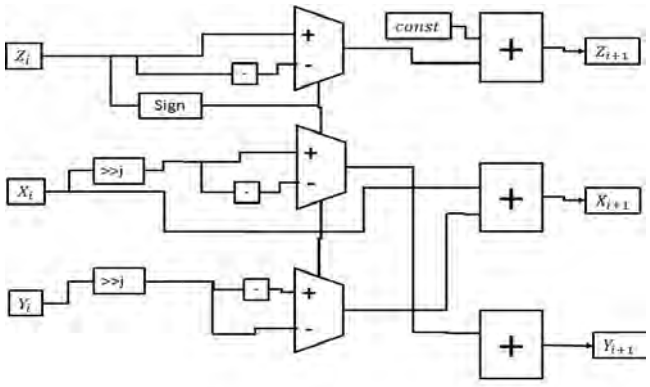


Fig. 8. Stage 4 and 5 block diagrams

Stage 6 (last stage): We are left with a residual angle gap, whose range is 0.875. For this reason, the rotator takes advantage of a nanorotator approximation with a non-constant, adaptive scaling coefficient: R = 1+ (S * 0.001953125i) for variadic phase, where S ∈ [0,8]. Considering the allowed values of S, the range of convergence is (-0.895, 0.895). The hardware compatible version of the coefficient is given in equation (19) and its architecture is shown in Figure 9. The stage 6 implementation requires extra logic: a scale decoder and an attenuation block.

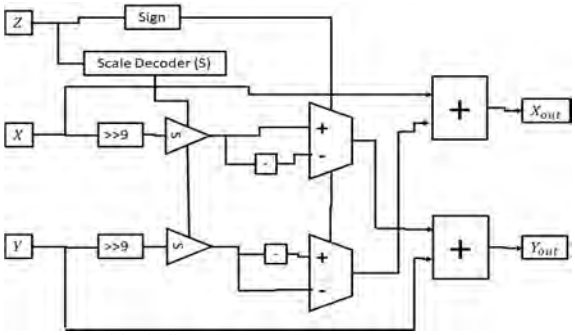


Fig. 9. Stage 6's block diagram

$$\begin{bmatrix} x_5 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & -d * 2^{-9} * S \\ d * 2^{-9} * S & 1 \end{bmatrix} \begin{bmatrix} x_4 \\ y_4 \end{bmatrix} \quad (19)$$

The scale decoder block determines the magnitude of the adaptive coefficient of S using the remaining phase, to make the residual of Z as close to 0 as possible. 9 combinations of S are obtained, see Table 2.

Table 2. Remaining angle range for S

Range	S	Range	S
(0, 0.0988]	0	(0. 4998, 0.5987]	5
(0. 0988, 0.1977]	1	(0. 5987, 0.6976]	6
(0. 1977, 0.2966]	2	(0. 6976, 0.7965]	7
(0. 2966, 0.3955]	3	(0.7965, 0.895]	8
(0. 3955, 0.4998]	4		

The attenuation block in Figure 9, depicted as a triangular block with "S" tag, multiplies the adaptive scale S to the shifted coordinate variables X and Y as defined in equation 19. Here, we use a regular multiplier.

The FPGA implementation integrates all these stages in series to complete the design. Given the combination of coefficients of all stages, the cumulative scale factor is K = 1.5757. Hence, we modify the initial vector in stage 1 by pre-dividing the values by this scale factor, which eliminates the other extra multipliers/dividers.

IV. RESULTS

We evaluate the design's performance by measuring the latency, resource usage, logic operator utilization, SFDR, and maximum frequency. These parameters provide trade-off considerations for a target application with specific requirements. For the sake of comparison, we provide values for three different CORDIC-based DDS implementations with and without a pipeline in the architecture. These are conventional CORDIC, modified Scaling Free CORDIC [3], and our proposed design.

Table 3 shows resource utilization based on the number of LUTs, FF, and RAM for a given target device. Here, the LUTs are the slice logic which is not necessarily using memory. We use ROM as memory: here its usage is measured in bits. In the table, CORDIC represents the conventional CORDIC (it stores 15 phases for the angle set and assumes 16 bits of variable's width). SF-CORDIC stands for modified Scaling Free CORDIC algorithm. The "P" next to the algorithm's name indicates the pipelined version. The initiation interval of every pipelined algorithm is 1, meaning the module can take input every clock cycle with no extra delay.

Table 3. Resource utilization

Algorithm	LUT	FF	ROM
CORDIC	764	84	240
CORDIC P	2506	840	
SF-CORDIC	479	98	96
SF-CORDIC P	835	340	
Proposed	400	140	
Proposed P	498	206	

Table 4 presents the utilization of logic operators. Mult, Add, Comp, Mux, and Shift stand for the multiplier, adder, comparator, multiplexer, and logic shifter. All these logic operators run with variables of 16 bits. The logic shifter accepts the variadic length of shift argument.

Table 4. Logic operator usage

Algorithm	Mult	Add	Register	Comp	Mux	Shift
CORDIC	1	7	5	4	21	2
CORDIC P	1	100	65	19	60	
SF-CORDIC		10	6	14	58	4
SF-CORDIC P		28	63	19	49	
Proposed	2	16	31	20	42	
Proposed P	2	24	68	22	32	

In Table 5, the values of SFDR are specified in dB. We compute the SFDR by fetching the results obtained from ILA signal analyzer in the MATLAB platform. Iteration in Table 5 indicates the number of rotations, this number is equal to the number of phases in the set. Latency is specified in the number of clock cycles. It represents the overall delay due to iterations and additional strategies such as domain folding and argument reduction techniques.

Table 6 lists the maximum frequency in MHz if implemented of a Xilinx Virtex-6 FPGA. The first column shows the maximum frequency for State-Machine (SM) based DDS and the second one lists that of the pipelined version.

Table 5. SFDR, iteration and overall latency

Algorithm	SFDR	iteration	Latency
CORDIC	92.7394	15	17
SF-CORDIC	56.8218	6	9
Proposed	72.2068	5	6

Table 6. Maximum Frequency

Algorithm	Max frequency	Max Frequency P
CORDIC	180	240
SF-CORDIC	226	354
Proposed	211	251

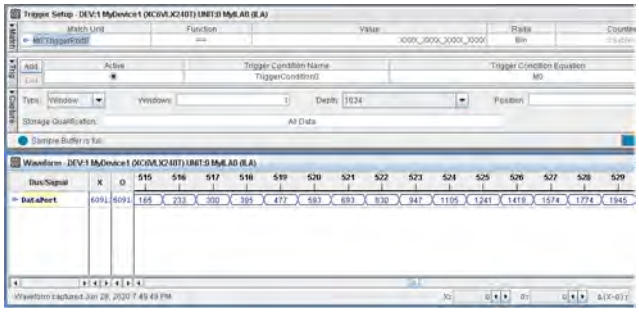


Fig. 10. ILA analyzer

In terms of resource utilization proposed design is better than the other designs both in pipelined and SM versions. Conventional CORDIC occupies the largest memory usage due to numerous angles in the set. Although it uses no memory for the pipelined version, the high number of iterations makes the increment of the other resources enormous. Memory is no longer needed because every stage is implemented separately and is active simultaneously, hence each stage is pre-assigned with a constant angle. SM based SF-CORDIC has the lowest resource as compared to our design, but its pipeline-based is behind the proposed design since it employs more iterations than ours. Our SM based design virtually doesn't need iteration because each stage employs a different type of rotator, which makes the resource usage close to that of pipeline-based design. To no surprise, our design occupies the smallest area and provides an energy consumption advantage.

SM based conventional CORDIC has the least overall logic operators, while the proposed design compares positively to the pipelined SF-CORDIC. The pipelined version of conventional CORDIC uses significantly more logic operators compared to the SM based one because 15 iterations that run on a set of resources are expanded into 15 identical sets of resources. The same explanation holds for the close figures of logic operator usage in the SM based and pipelined version of the proposed design. Here, the synthesize tool optimizes the resource allocation by substituting shifter by a concatenation operator due to constant bit shifting. Consequently, the number of logic shifters may not be the same as shown in the block diagrams.

Low latency is desired to enhance the throughput and efficiency of the communication system because delay in the system slows down quantum feedback - a bottleneck and great challenge for quantum-enhanced communication systems. With a latency of just 6 clock cycles, the proposed design is superior to the other algorithms. Although the number of iterations of SF-CORDIC is very close to that of the proposed design, there is an extra delay of 3 clock cycles due to a required additional compensation to the rotation.

The SF-CORDIC has the highest maximum frequency due to one-directional rotation. Our design has a moderate maximum frequency for its implementation. The conventional CORDIC achieves the highest SFDR value due to the high number iteration.

Our design achieves moderate SFDR yet the lowest latency, with approximately 20 dB SFDR and 64% latency reductions compared to that of the conventional CORDIC design.

V. CONCLUSION

In conclusion, we report a new memory free low latency DDS architecture. Generally, complex value computations could be employed to calculate the trigonometric equation, but those calculations are computationally difficult and energy inefficient. To avoid calculation-related inefficiencies, the common approach is to use a LUT with phase being the input and amplitude being the output. To generate a desired smooth radio-frequency signal small-step quantization is needed, requiring a larger LUT. The LUT requirement leads to an increase in memory usage and may lead to the reduction of the maximum frequency of the FPGA design which would also limit modulation capabilities. On the other hand, CORDIC technique offers low complexity and memory-free trigonometric calculation approach, with the expense of extra latencies to complete the computation. We use a pipelined approach to shorten latency even more. Thus, in our design, the sinusoidal wave amplitude is obtained every cycle, thus maximizing our modulation capabilities. To make a quantum measurement enhanced transceiver, we choose the modulation scheme which includes choosing the number of states M, the frequency, and the initial phase detuning between the adjacent states and other communication parameters. All M states are being prepared in parallel at all times, and the active output state is picked according to the encoding and measurement protocols. Because M could be quite large (up to 16 in our implementation) the low-resource usage DDSs are essential for this purpose. In communication links, sensitivity is often measured as the probability to receive an erroneous symbol with certain energy at the receiver. Classical receivers have a sensitivity limit known as the standard quantum limit (SQL). This limit arises from the inevitable shot noise on the idealized classical receiver scheme - a homodyne measurement followed by a perfect detector with no noise of its own and with the 100% detection-efficiency. The SQL is accessible only through quantum measurement. With the help of the described DDS, we have implemented a quantum-measurement telecommunication testbed and demonstrated that the sensitivity of a telecommunication channel is better than SQL for many different modulation protocols, including quantum-measurement specific modulation protocols, described elsewhere [10].

A specific target application for this DDS is to build a quantum measurement enhanced transceiver. Particularly, we intend to use modulation schemes that require a simultaneous phase and frequency modulation. Our novel design achieves the shortest latencies, maximizes modulation capabilities, and uses the minimal footprint compared to other CORDIC-based DDSs.

ACKNOWLEDGMENT

This work is partially supported by National Science Foundation USA through ECCS 1927674.

REFERENCES

[1] I.A. Burenkov, M.V. Jabir, N.F.R. Annafianto, A. Battou, and S.V. Polyakov, "Experimental Demonstration of Time Resolving Quantum Receiver for Bandwidth and Power Efficient Communications", Proc. of *CLEO Conf. on Laser Science to Photonic Applications*, California, USA, 2020.

[2] P. Saravanan and S. Ramasamy, "Sine/cos generator for direct digital frequency synthesizer using pipelined CORDIC processor," Proc. of *Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1-6, Tiruchengode, 2013.

[3] R. Xin, X. Zhang, H. Li, Q. Wang, and Z. Li, "An Area Optimized Direct Digital Frequency Synthesizer Based on Improved Hybrid CORDIC Algorithm," Proc. of *Int. Workshop on Signal Design and Its Applications in Communications*, pp. 243-246, Chengdu, China, 2007.

[4] Y. Xue and Z. Ma, "Design and Implementation of an Efficient Modified CORDIC Algorithm," Proc. of *IEEE Int. Conf. on Signal and Image Processing (ICSIP)*, pp. 480-484, Wuxi, China, 2019.

[5] M.M. Anas, R.S. Padiyar, and A.S. Boban, "Implementation of Cordic Algorithm and Design of High Speed Cordic Algorithm," Proc. of *Int. Conf. on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 1278-1281, Chennai, India, 2017.

[6] Y.S. Gener, S. Gören, and H.F. Ugurdag, "Lossless Look-Up Table Compression for Hardware Implementation of Transcendental Functions," Proc. of *IFIP/IEEE Int. Conf. on Very Large Scale Integration (VLSI-SoC)*, pp. 52-57, Cuzco, Peru, 2019.

[7] W. Shuqin, H. Yiding, Z. Kaihong, and Y. Zongguang, "A 200MHz Low-Power Direct Digital Frequency Synthesizer Based on Mixed Structure of Angle Rotation," Proc. of *IEEE Int. Conf. on ASIC*, pp. 1177-1179, Changsha, China, 2009.

[8] K. Maharatna, S. Banerjee, E. Grass, M. Krstic, and A. Troya, "Modified Virtually Scaling-Free Adaptive CORDIC Rotator Algorithm and Architecture," in *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, pp. 1463-1474, 2005.

[9] M. Garrido, P. Källström, M. Kumm, and O. Gustafsson, "CORDIC II: A New Improved CORDIC Algorithm," in *IEEE Tran. on Circuits and Systems II: Express Briefs*, vol. 63, pp. 186-190, 2016.

[10] I.A. Burenkov, O.V. Tikhonova, and S.V. Polyakov, "Quantum Receiver for Large Alphabet Communication," *Optica*, vol. 5, pp. 227-232, 201

Features of JFET Computer Models in Microcurrent Mode on Exposure to Low Temperatures and Neutron Fluence

Oleg Dvornikov
JSC "MNIPI",
Minsk city, Belarus
oleg_dvornikov@tut.by

Valentin Dziatlau
INP BSU,
Minsk city, Belarus
nitnelaff@gmail.com

Vladimir Tchekhovski
INP BSU,
Minsk city, Belarus
vtchek@hep.by

Nikolay Prokopenko
Member, IEEE
DSTU, IPPM RAS,
Rostov-on-Don city, Zelenograd city, Russia
prokopenko@sssu.ru

Anna V. Bugakova
Student Member of IEEE
DSTU,
Rostov-on-Don city, Russia
annabugakova.1992@gmail.com

***Abstract*—The current-voltage curves (CVCs) of silicon p-JFET and n-JFET, measured in the range of drain currents of about 60 dB, are considered. To describe the CVCs of the JFET in micromode, that is for the drain currents less than 1 μA, it is proposed to use the Shichman-Hodges model built into Spice-like programs with a set of parameters identified in the micromode. The modeling results and measurement data of the CVCs in the micromode is shown, as well as the simulation results at temperatures up to minus 197°C and on exposure to neutron fluence are presented. The adequacy of the p-JFET models is set accurate for the design of analog circuits. It was difficult to separate the linear area from the saturation area at the output CVC in a circuit with a common n-JFET source at drain currents less than 1 μA. To solve the problem with this CVC, we revealed the greatest discrepancy between measurements and modeling. Therefore, it is permissible to use the identified parameters of the n-JFET model only for evaluative modeling of microcircuits in the micro mode.**

Keywords—current-voltage curves, JFETs, JFET models, cryogenic electronics, nuclear hardness, analog sensor interfaces

I. INTRODUCTION

The junction gate field-effect transistors (JFETs) are often used in electronic equipment to provide low low-frequency noise. In low-noise analog integrated circuits (ICs), JFETs are usually input transistors, and metal-oxide-semiconductor field-effect transistor (MOSFET) or bipolar junction transistors (BJT) are used as the remaining active elements [1-11], and input JFETs can be either with a p-type channel (p-JFET) or n-type channel (n-JFET).

In recent years, the interest in the use of the JFETs in space equipment has grown significantly, due to their high nuclear hardness and the preservation of characteristics at low temperatures [12,13], up to the temperature of liquid nitrogen.

Due to their good low-temperature characteristics, the JFETs are used in analog sensors in which the input JFET is significantly cooled to reduce thermal noise [14,15]. In such devices, it is necessary to transmit a signal from the cooled input JFET via cable to the equipment that is in normal conditions. To exclude self-heating of the cooled unit, it is desirable that it contains only the input JFET and a readout circuit operating in low power consumption mode. Ensuring the required parameters of the readout circuit at low temperatures is most easily achieved with the use of complementary JFETs – p-JFET and n-JFET with approximately the same cut-off voltage [13].

It is understood that accelerated design of analog devices is possible only when using circuit simulation. However, the Shichman-Hodges model used in many Spice-like programs does not quite adequately characterize the current-voltage curves (CVCs) of the JFET, especially in the micromode and during the transition from the linear area of the CVC to saturation. In connection with the mentioned information by various experts, the work has been done to improve the JFET models, including the description of the CVCs at extremely low drain currents [16-18]. Some of the created models are made in the Verilog-A language for compatibility with Spice-like programs [19,20]. Unfortunately, these models do not take into account the influence of penetrating radiation dose and cryogenic temperatures.

Earlier, we modernized the Shichman-Hodges model, which made it possible, with sufficient accuracy for many cases, to simulate the static parameters of analog ICs based on complementary JFETs at low temperatures, up to minus 197°C, and on exposure to neutron fluence up to 10^{15} n/cm^2 [21]. In addition, we experimentally established that the effect of ^{60}Co gamma rays with an absorbed dose of up to 1 Mrad does not lead to a significant change in the CVC of the JFET.

The purpose and objective of this paper is to experimentally study the CVCs of complementary JFETs manufactured by JSC "Integral" and to search for technical solutions that simulate the CVCs of JFET in a wide range of drain currents, at temperatures up to minus 197°C and on exposure to neutron fluence.

The study has been carried out at the expense of the grant from the Russian Science Foundation (Project No. 16-19-00122-P).

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

II. Measurement Technique and Research Samples

A study was conducted of test JFETs that were manufactured at JSC "Integral" on the "Inch-R/NJFET" and "Inch-R/PJFET" processing sequences [13]. Semiconductor substrates of the same type, epitaxial layers, interconnects and almost identical p- and n-type semiconductor layers with depth X_J and sheet resistance R_S are used in both processing sequences. R_S p+ "hidden" layer, conductivity type, X_J and R_S channel semiconductor area and n-JFET and p-JFET gates are differences between "Inch-R/NJFET" and "Inch-R/PJFET" processing sequences. The impurity concentration profile has a nonmonotonic nature in the channel of both JFETs. The thickness of the conductive part of the channel in the absence of external voltage will be equal to 0.3 μm for p-JFET, and 3.4 μm for n-JFET, and the average impurity concentration has the following characteristics for p-JFET is $1 \cdot 10^{17}$ cm^{-3}, for n-JFET it is $6.3 \cdot 10^{14}$ cm^{-3}. In topology, the ratio of the gate width of the transistor to its W/L length is 50 μm/6 μm for p-JFET and 260 μm/6 μm for n-JFET.

JFET measurements were carried out automatically using the IPPP-1 instrument [22]. We obtained the characteristics of the output current-voltage curve, which is determined by the drain current I_D dependence on the drain-source voltage V_{DS} for several gate-source voltage V_{GS}, as well as the transfer current-voltage curve, here the dependence of I_D on V_{GS} for fixed values of V_{DS}. The obtained results were consistent with the Shikhman – Hodges model:

- in the saturation area for $V_{SD} \geq V_{TH} - V_{GS}, V_{GS} < V_{TH}$ (the signs are given for p-JFET)

$$I_D = \beta(1 + \lambda V_{SD})(V_{TH} - V_{GS})^2, \quad (1)$$

- in the linear area for $0 < V_{SD} < V_{TH} - V_{GS}$, $V_{GS} < V_{TH}$

$$I_D = \beta(1 + \lambda V_{SD}) V_{SD}[2(V_{TH} - V_{GS}) - V_{SD}], \quad (2)$$

where β is specific transconductance (β~W/L); λ is coefficient of the channel length modulation; V_{TH} is cut-off voltage (for p-JFET – a positive value).

III. Results of Measurement

5 samples of n-and p-type JFET were measured, then collected in metal-ceramic packages. Processing of the measurement results was carried out as follows:

- the dependences of I_D on V_{GS} were constructed for all samples at $|V_{DS}| = 5$ V;
- a "typical" sample having the CVC closest to the average one was visually determined;
- according to previously developed methods [23], for a "typical" sample, the model parameters V_{TH}, β, λ (Table I) were identified in the CVC saturation area, and dependences $\sqrt{I_D}$ on V_{GS}, I_D/I_{DMAX} on V_{DS}, where $I_{DMAX} = I_D$ at V_{GS}=const, $|V_{DS}|$=10 V were constructed.

TABLE I. The Results of the Parameter Identification of the Shichman-Hodges Model

Type of transistor	Range $\|I_D\|$	λ^{-1}, V	$\|V_{TH}\|$, V	β, μA /V^2
p-JFET	>10 μA	79.4-91.1	1.753	53.29
	1 μA - 10 μA	49.1-79.4	1.726	53.25
	0.1 μA - 1 μA	45.9-49.1	1.804	21.16
n-JFET	>15 μA	21.9-45.5	1.093	420.25
	1 μA - 10 μA	14.5-19.9	1.158	176.89
	0.2 μA - 1 μA	12.1-14.4	1.295	19.36

Since during all measurements the JFET sources were connected to the no-volt bus, the value and the polarity of the voltage supplied to the gate and drain output are indicated in the figures. Note that the inverse value of the parameter λ, also known as the Earley voltage V_A=λ-1, allows you to compare the output low-signal resistance of bipolar transistors and JFET (see Table 1).

The data obtained allow us to formulate the conclusions about the specificity of the JFET CVCs manufactured by JSC "Integral":

1. From (1) it follows that the dependence of $\sqrt{I_D}$ on V_G at V_D=const should be a straight line. However, in Fig. 1, Fig. 2 at low drain currents, approximately less than 1 μA, a deviation of the dependence of $\sqrt{I_D}$ on V_G from a straight line is observed, i.e. the CVC of the studied transistors at low drain currents does not correspond to the Shichman-Hodges model.

2. On the relationships of the normalized drain current I_D/I_{DMAX} on V_D, shown in Fig. 3, Fig. 4, curves 1 refer to the area of high currents, and curves 2 to the micromode. For the p-JFETs, both at high and low currents, a linear area and a saturation area exist at the output CVC. For the n-JFETs at high currents, a very smooth transition from the linear area to the saturation area is observed, and in the micromode on the CVC it is difficult to separate the linear area from the saturation area. Thus, the application of the Shichman-Hodges model to describe the CVCs of the n-JFET at low currents will give a large error.

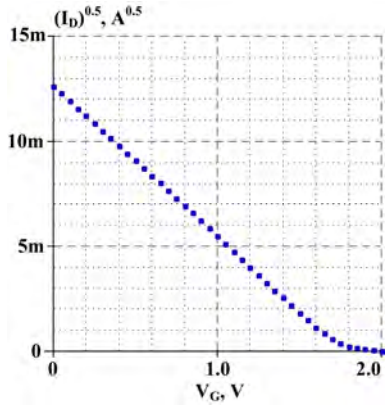


Fig. 1. The obtained measurement data of the relationship of $\sqrt{I_D}$ on V_G for p-JFET at V_D=-5 V.

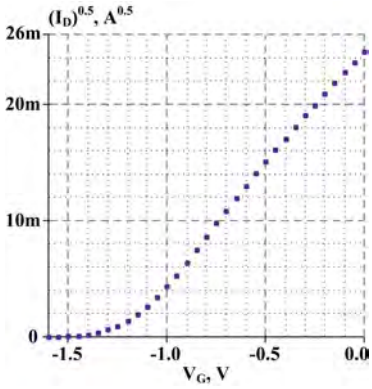


Fig. 2. The obtained measurement data of the relationship of $\sqrt{I_D}$ on V_G for n-JFET at V_D=5 V.

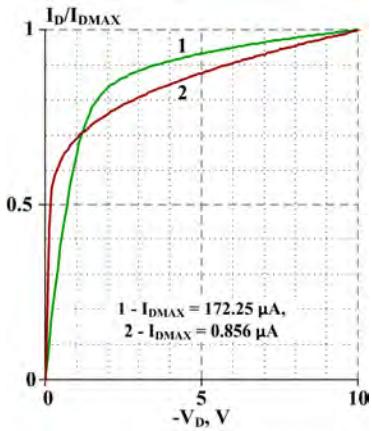


Fig. 3. The obtained measurement data of the relationship of the normalized drain current I_D/I_{DMAX} on V_D for p-JFET.

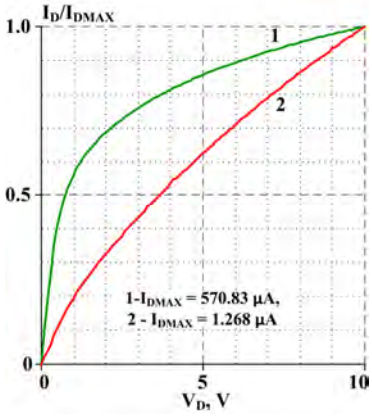


Fig. 4. The obtained measurement data of the relationship of the normalized drain current I_D/I_{DMAX} on V_D for n-JFET.

3. The final conclusion about the applicability of the Shichman-Hodges model for describing the JFETs CVCs manufactured at JSC "Integral" can be made on the basis of the model parameters contained in Table I, namely:

- for the JFETs, it is advisable to use two sets of model parameters: the first one – for the drain current range of 0.1 µA-1 µA, the second one – for the drain current of more than 10 µA;
- in the range of drain currents of 1 µA-10 µA, the simulating error will be maximum, and the greatest discrepancy between the modeling results and measurement data should be expected from the output CVC of the n-JFET due to the strong dependence $\lambda = \lambda(I_D)$.

Since the most interesting is the assessment of the possibility of using the Shichman-Hodges model to describe the operation of the JFET in micromode, we simulated the CVCs at low currents using the parameters of Table I for the range of 0.1 µA-1 µA and comparison of the obtained results with the measurements. The dependences shown in Fig. 5 - Fig. 8 allow us to state that sufficient adequacy of the Shichman-Hodges model is provided only in a limited range of small drain currents.

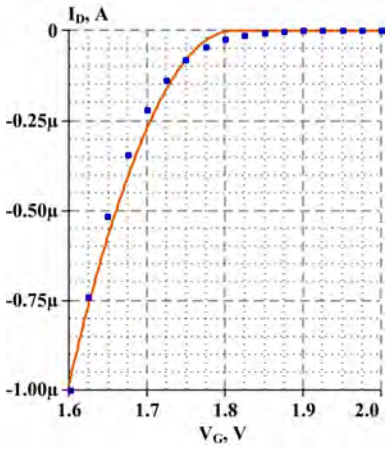


Fig. 5. The obtained measurement data (points) and modeling (solid line) of the relationship of I_D on V_G for the p-JFET at V_D = -5 V.

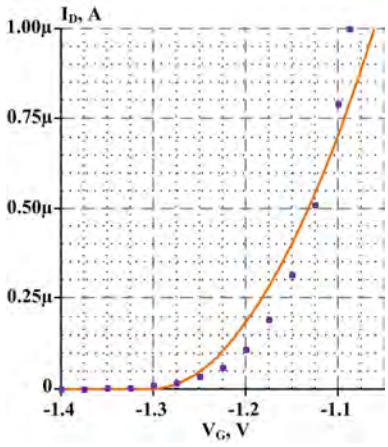


Fig. 6. The obtained measurement data (points) and modeling (solid line) of the relationship of I_D on V_G for the n-JFET at V_D = 5 V.

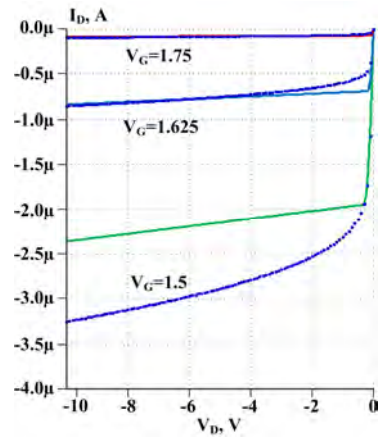


Fig. 7. The obtained measurement data (points) and modeling (solid line) of the relationship of I_D on V_D for the p-JFET.

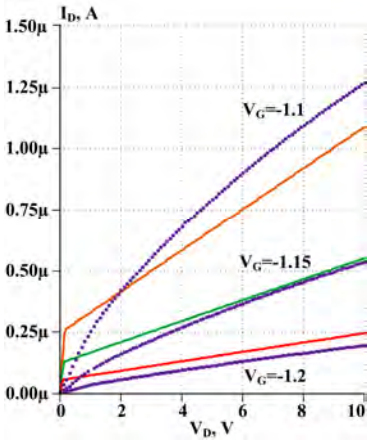


Fig. 8. The obtained measurement data (points) and modeling (solid line) of the relationship of I_D on V_D for the n-JFET.

As noted earlier, there are a number of JFET models that allow high-precision simulation of the output and transfer CVCs in a wide range of drain currents. The application of the Shichman-Hodges model requires the use of two sets of model parameters, makes it difficult to obtain reliable simulation results in a certain transition area of drain currents (for the studied transistors at 1 μA-10 μA), and therefore the appropriateness of using this model is doubtful.

From our point of view, the main reason for using the Shichman-Hodges model is that it makes possible in a simple way to take into account the effect of cryogenic temperatures up to minus 197°C and the effect of neutron fluence on the CVCs [21].

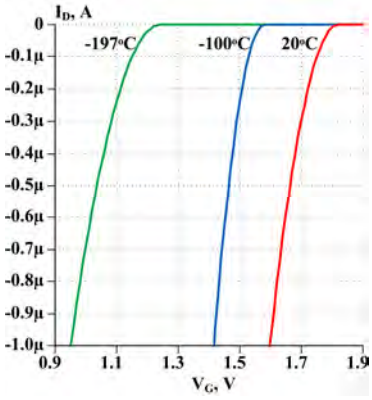


Fig. 9. The computer modeling of the relationship of I_D on V_G for the p-JFET at V_D = -5 V and three temperature values.

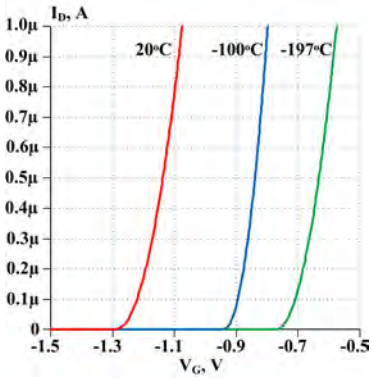


Fig. 10. The computer modeling of the relationship of I_D on V_G for the n-JFET at V_D = 5 V and three temperature values.

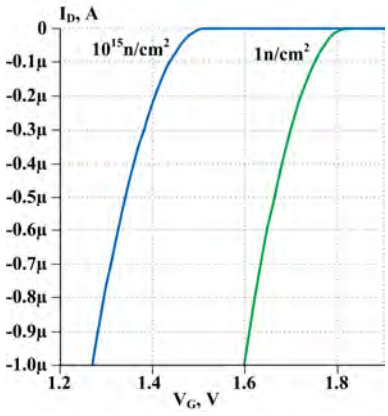


Fig. 11. The computer modeling of the relationship of I_D on V_G for the p-JFET at V_D = -5 V and different neutron fluences.

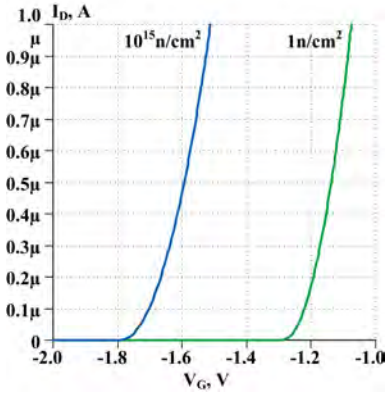


Fig. 12. The computer modeling of the relationship of I_D on V_G for the n-JFET at V_D = 5 V and different neutron fluences.

Taking into account the previously obtained dependences of the temperature and radiation changes in the JFET parameters [21], the transfer CVC was simulated in micromode at different temperatures and neutron fluences (1 n/cm^2 corresponds to the normal condition before irradiation).

The simulation results shown in Fig. 9-Fig. 12, correspond to the basic experimental data for the JFET manufactured by JSC "Integral":

- decrease of $|V_{TH}|$ with decreasing temperature;
- nonmonotonic change of transconductance from the temperature with maximum value of transconductance about T out °C;
- a qualitative difference in the change in the CVCs of the JFET with a different type of channel conductivity upon irradiation with neutrons, namely, with increasing the neutron fluence $|V_{TH}|$ and $|I_D|$ for the n-JFET- increase, and for the p-JFET decrease.

IV. Conclusion

We measured the output and transfer CVCs at drain currents from hundreds of nanoamperes to hundreds of microamperes for silicon n-JFET and p-JFET by JSC "Integral". At drain currents of approximately less than 1 μA, the deviation of the dependence of $\sqrt{I_D}$ from V_{GS} from the straight line on the transfer CVC for both types of JFETs, the absence of a clearly defined linear area and the saturation area at the output CVC of n-JFET were established.

It is concluded that the application of the Shichman-Hodges model with a set of parameters identified in micromode enables us to obtain the sufficient adequacy of the CVC simulation in a limited range of low drain currents.

Despite the inherent drawbacks, the application of the Shichman-Hodges makes possible to take into account the effect of cryogenic temperatures and neutron fluence on the JFET CVCs in micromode.

REFERENCES

[1] D. A. Fleischer, S. Shekar, S. Dai, R. M. Field, J. Lary, J. K. Rosenstein, K. L. Shepard, "CMOS-Integrated Low-Noise Junction Field-Effect Transistors for Bioelectronic Applications," in *IEEE Electron Device Letters*, vol. 39, no. 7, pp. 931-934, July 2018. DOI: 10.1109/LED.2018.2844545.

[2] K. Nidhi and M. Ker, "A CMOS-Process-Compatible Low-Voltage Junction-FET With Adjustable Pinch-Off Voltage," in *IEEE Transactions on Electron Devices*, vol. 64, no. 7, pp. 2812-2819, July 2017. doi: 10.1109/TED.2017.2706423.

[3] Y. Shi, R. M. Rassel, R. A. Phelps, P. Candra, D. B. Hershberger, X. Tian, S. L. Sweeney, J. Rascoe, B. Rainey, J. Dunn, D. Harame, "A cost-competitive high performance Junction-FET (JFET) in CMOS process for RF & analog applications," *2010 IEEE Radio Frequency Integrated Circuits Symposium*, Anaheim, CA, 2010, pp. 237-240. DOI: 10.1109/RFIC.2010.5477348.

[4] T. Yang, J. Lu and J. Holleman, "A high input impedance low-noise instrumentation amplifier with JFET input," *2013 IEEE 56th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Columbus, OH, 2013, pp. 173-176. DOI: 10.1109/MWSCAS.2013.6674613.

[5] K. Y. J. Hsu and T. Chuang, "An input buffer with monolithic JFET in standard BCD technology for sensor applications," *2015 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*, Singapore, 2015, pp. 784-787. DOI: 10.1109/EDSSC.2015.7285235.

[6] Z. He, C. Wang, G. Fan, Y. Zhou and Y. Yang, "Design of A High Input Impedance OPA with BiJFET Technology," *2019 IEEE 2nd International Conference on Electronics Technology (ICET)*, Chengdu, China, 2019, pp. 233-236. DOI: 10.1109/ELTECH.2019.8839538.

[7] 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)*, Dresden, 2018, pp. 290-293. DOI: 10.1109/ESSCIRC.2018.8494262.

[8] T. Yang, R. Huang and S. Bai, "Piecewise Linear Approximation for Extraction of JFET Resistance in SiC MOSFET," in *IEEE Transactions on Electron Devices*, vol. 65, no. 10, pp. 4455-4461, Oct. 2018. doi: 10.1109/TED.2018.2862460.

[9] K. Xia, C. C. McAndrew and B. Grote, "Dual-Gate JFET Modeling I: Generalization to Include MOS Gates and Efficient Method to Calculate Drain–Source Saturation Voltage," in *IEEE Transactions on Electron Devices*, vol. 63, no. 4, pp. 1408-1415, April 2016. doi: 10.1109/TED.2016.2525737.

[10] N. Makris, F. Jazaeri, J. Sallese and M. Bucher, "Charge-Based Modeling of Long-Channel Symmetric Double-Gate Junction FETs—Part II: Total Charges and Transcapacitances," in *IEEE Transactions on Electron Devices*, vol. 65, no. 7, pp. 2751-2756, July 2018. doi: 10.1109/TED.2018.2838090.

[11] O. V. Dvornikov, V. L. Dziatlau, V. A. Tchekhovski, N. N. Prokopenko and A. V. Bugakova, "BiJFet Array Chip MH2XA030 — a Design Tool for Radiation-Hardened and Cryogenic Analog Integrated Circuits," *2018 IEEE International Conference on Electrical Engineering and Photonics (EExPolytech)*, St. Petersburg, 2018, pp. 13-17. DOI: 10.1109/EExPolytech.2018.8564415.

[12] O. V. Dvornikov, V. L. Dziatlau, 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 IEEE SIBCON*, Astana, 2017, pp. 1-6. DOI: 10.1109/SIBCON.2017.7998507.

[13] O. V. Dvornikov, N. N. Prokopenko, A. V. Bugakova, V. A. Tchekhovski and I. V. Maliy, "Cryogenic Operational Amplifier on Complementary JFETs," *2018 IEEE East-West Design & Test Symposium (EWDTS)*, Kazan, 2018, pp. 1-5. DOI: 10.1109/EWDTS.2018.8524640.

[14] A. Pullia, F. Zocca, S. Riboldi, D. Budjas, A. D'Andragora and C. Cattadori, "Cryogenic Performance of a Low-Noise JFET-CMOS Preamplifier for HPGe Detectors," in *IEEE TNS*, vol. 57, no. 2, pp. 737-742, April 2010. DOI: 10.1109/TNS.2009.2038697.

[15] C. Boiano, R. Bassini, A. Pullia and A. Pagano, "Wide-dynamic-range fast preamplifier for pulse shape analysis of signals from high-capacitance detectors," in *IEEE TNS*, vol. 51, no. 5, pp. 1931-1935, Oct. 2004. DOI: 10.1109/TNS.2004.832308.

[16] V.N. Biryukov, "Template modeling of a p-channel MOSFET. Zhurnal Radioelektroniki," *Journal of Radio Electronics*, 2019, No. 2. DOI: 10.30898/1684-1719.2019.2.11.

[17] N. Makris, F. Jazaeri, J. Sallese, R. K. Sharma and M. Bucher, "Charge-Based Modeling of Long Channel Symmetric Double-Gate Junction FETs—Part I: Drain Current and Transconductances," in *IEEE TED*, vol. 65, no. 7, pp. 2744-2750, July 2018. DOI: 10.1109/TED.2018.2838101.

[18] N. Makris, M. Bucher, F. Jazaeri and J. Sallese, "CJM: A Compact Model for Double-Gate Junction FETs," in *IEEE Journal of the Electron Devices Society*, pp. (99):1-9, October 2019. DOI: 10.1109/JEDS.2019.2944817.

[19] N. Makris, M. Bucher, F. Jazaeri and J. Sallese, "A Compact Model for Static and Dynamic Operation of Symmetric Double-Gate Junction FETs," *2018 48th European Solid-State Device Research Conference (ESSDERC)*, Dresden, 2018, pp. 238-241. DOI: 10.1109/ESSDERC.2018.8486848.

[20] K. Xia and C. C. McAndrew, "JFETIDG: A Compact Model for Independent Dual-Gate JFETs With Junction or MOS Gates," in *IEEE TED*, vol. 65, no. 2, pp. 747-755, Feb. 2018. DOI: 10.1109/TED.2017.2786043.

[21] O. Dvornikov, V. Dziatlau, V. Tchekhovski, N. Prokopenko, A. Zhuk and A. Bugakova, "Modernization of Low-Temperature JFET Models Built into LTspice CAD Systems, Taking into Account the Results of their Experimental Study," 2020 IEEE Latin America Electron Devices Conference (LAEDC), San José, Costa Rica, Feb. 25 - 28, 2020, pp. 1-4.

[22] B. N. Lisenkov, N. V. Gritsev, A. G. Petrovich, "Control of electrical parameters of electronic components" [Online]. Available: http://mnipi.by/articles/kontrol-elektricheskih-parametrov-elektronnyh-komponentov.html (In Russian)

[23] O. Dvornikov, Yu. Shulgevich, "Methods for identifying the parameters of integrated transistor models. Part 4. Identification of the parameters of the Shichman-Hodges model of field-effect transistors with a p-n junction," *J. Modern electronics*, 2009, No. 8. Pp. 50–57. (In Russian)

Synthesis of Approximate Combinational Circuits based on Logic Regression Approach

Alexander Stempkovsky
DSc, prof., scientific director at Institute for Design Problems in Microelectronics of RAS
Moscow, Russia
stal09@ippm.ru

Dmitry Telpukhov
Department of Integrated Circuits Design Methodology
Institute for Design Problems in Microelectronics of RAS
Moscow, Russia
dmtr@ippm.ru

Roman Solovyev
Department of Integrated Circuits Design Methodology
Institute for Design Problems in Microelectronics of RAS
Moscow, Russia
zf-turbo@yandex.ru

***Abstract*— Approximate synthesis is a modern trend in the field of logic synthesis, which makes it possible to obtain much more compact, high-speed and reliable solutions due to weakening requirements for the accuracy of the implemented functions. For a number of applications, small distortions at outputs can be more than acceptable, and the improvement of characteristics is a powerful argument in favor of this method. We propose a new approach to the approximate synthesis of combinational logic, which is built upon solving the logic regression problem with the use of iterative methods based on two-level minimization. In the paper we state the logic regression problem and describe its possible applications. Solution based on two-level minimization methods is presented. We performed experimental research, which demonstrates high efficiency of the method. We revealed a fundamental drawback of the proposed method, that is, complexity of implementing circuits with a large number of linearly non-separable elements (XOR, NXOR). The ways to overcome this drawback are outlined.**

Keywords—Approximate synthesis, logic regression, two-level minimization, ESPRESSO logic minimizer.

I. INTRODUCTION

The paper considers the problem of logic regression on high dimensional Boolean space. Traditionally, the problem was addressed for a wide range of tasks, including genetics or identification of predictors in medical data [1]. Current research suggests the use of this methodology in the context of microelectronics [2]. In this area, logic regression can be applied to the following tasks:

- equivalence check with a compact semantic expression output;
- functional circuit correction (engineering change order – ECO): differences/patch generation;
- model validation;
- logic synthesis.

With reference to synthesis problems [3-4], due to logic regression methodology, we can turn to the popular modern line of research related to the approximate synthesis of logic circuits [5]. In this framework, it is suggested to use some flexibility resulting from the feasibility of certain deviations from nominal output values, aimed at reducing area and delay of the circuit, as well as increasing fault tolerance [6] and yield percentage. Error at the outputs of logic circuits may be acceptable for a wide range of applications, including audio, video, graphics, and wireless communications, in case when failures are non-critical and the probability of their occurrence does not exceed the specified thresholds.

II. STATEMENT OF THE PROBLEM AND PREVIOUS WORK

In this work, the task is to restore a logic function and build the smallest approximate logic circuit from a known subset of its values at certain points.

The statement of the problem implies that there is a certain black box that can produce values at the outputs when certain values are fed into inputs. We have to restore combinational circuit from input-output pairs obtained from the black box (Fig. 1) [7].

The trivial solution is exhaustive enumeration of all possible inputs and constructing a circuit for the obtained truth table. Further logic synthesis could be carried out using both traditional (DNF, BDD) and new synthesis approaches [8]. However, for functions with more than 30 inputs, this method becomes inapplicable, and the development of more complex approaches is required.

Efficient solution to the problem opens up broad prospects for logic circuits synthesis based solely on the input vectors and responses of the system. This is relevant for a wide range of tasks, starting with engineering change order (ECO) and restoring circuits from ready-made chips, ending with the automatic translation of any high-level description into a logic circuit.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

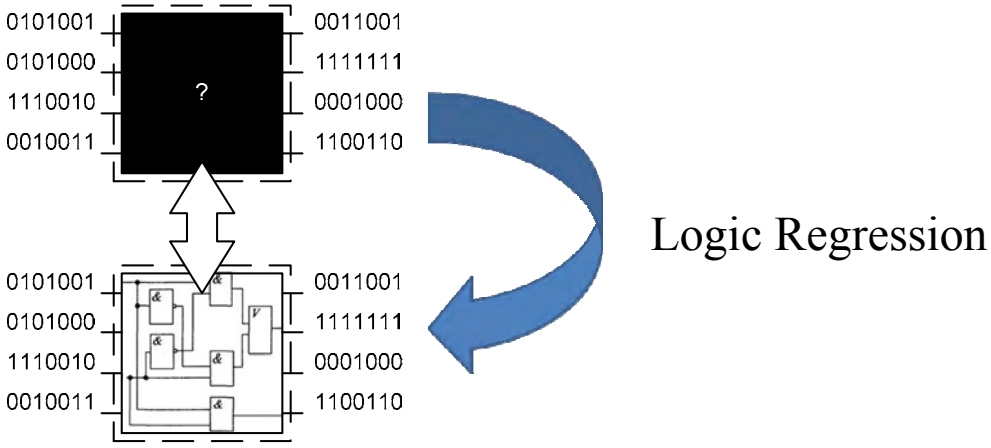


Fig. 1. Formulation of logic regression problem

The problem of logic regression was most developed in the field of bioinformatics [9][10]. The mentioned papers describe an approach related to manipulating logic trees based on the simulated annealing algorithm.

The stated problem can also be solved through the use of genetic algorithms [11] and Cartesian genetic programming methods [12]. Within the framework of these approaches, circuits can be evolved, and the proximity of the circuit to the black box function can be used as fitness function. Drawbacks of these methods are high computational load and relatively slow rate of convergence. As a consequence, they are inapplicable for medium and large circuits.

Solutions gained from another similar area can be applied to logic regression problem. This is the task of ECO (Engineering Change Order), and in particular, generating patches for functional correction of circuits. In [13], an approach was proposed for automated patch generation using conflict-based greedy method of finding a basis.

All existing methods have low scalability, and converge slowly in the case of large circuits. In this work, we propose a completely different approach not related to structural synthesis, but based on fast heuristic methods of two-level minimization.

III. ITERATIVE REGRESSION ALGORITHM BASED ON LOGIC SYNTHESIS METHODS

The paper proposes an iterative approach based on logic synthesis for regression of approximate logic circuits. The iterative method of logic regression can be implemented on the basis of standard two-level minimization, in particular, on the basis of the ESPRESSO algorithm [14]. We describe the contents of the proposed iterative approach (Figure 2).

The algorithm starts with generation of random input stimuli and corresponding output vectors. Further, we have to synthesize some implicant cover based on the output. At this step, we can either use existing methods and tools, or develop fundamentally new concepts. As a well-known solution, heuristic algorithm of two-level minimization ESPRESSO is well suited.

Assume that unknown values are at don't care state. The program will produce some compact disjunctive normal form (DNF) (or conjunctive normal form (CNF)) based on the known values of input and output signals, setting don't cares to 0 or 1. Such cover defines some initial solution, the quality of which is evaluated at the next step of the algorithm.

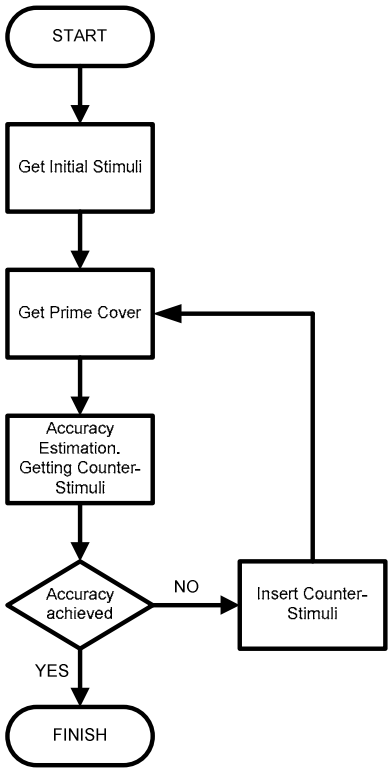


Fig. 2. Generalized algorithm for iterative logic regression method based on logic synthesis

We estimate accuracy by the Monte Carlo simulation method. Those stimuli that conflict with the newly set states of the current cover are written to the list of counter-stimuli. If the quantity of correct answers exceeds the specified threshold, the algorithm is completed. Otherwise, counter-stimuli are inserted. This procedure is defined as shown in Figure 3. A certain number of counter-stimuli are added to the list of implicants one by one, at the same time conflicting implicants are excluded from the list.

Then minimization algorithm is then launched again, taking into account remaining implicants and new counter-stimuli. At every iteration step, it forms a new cover that converges towards the best solution. This iterative process combines two basic assumptions. The first is that at each step minimization algorithm tries to find the minimum cover for the available data. Thus, the rule of Occam's razor is implemented – of all possible solutions, the simplest is chosen. The second basic assumption is that incorrect implicants are more often refuted and eliminated by counter-stimuli, while correct implicants are preserved from iteration to iteration – thus forming the basis for new assumptions.

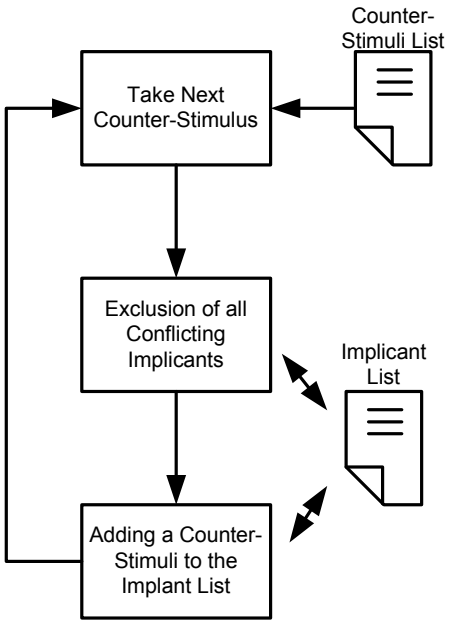


Fig. 3. Insertion of counter-stimuli to the list of implicants

Consider how the presented algorithm operates for a simple example of four-input functions described by Karnaugh maps (Figure 4). Let a black box implement a function of four variables. The algorithm starts with generation of first four initialization stimuli and output vectors. Next, the ESPRESSO algorithm works and minimizes the resulting function on the assumption that empty cells are don't cares. The algorithm works in the basis of DNF, therefore, it tries to find the smallest possible number of biggest areas of '1's. Further, by simulation, the algorithm detects a counter-stimulus (mismatch of the assumption to the black box) in zero area. Then it makes a new assumption – this time, incorrect. After that, it finds a new counter-stimulus in zero area and makes a valid hypothesis. At the final step, the counter-stimulus forces us to exclude the wrong area, thus, the resulting solution is equivalent to the reference function.

Consider the number of stimuli used: 4 initialization stimuli and 3 counter-stimuli; it turns out that the algorithm was able to restore the correct solution by analyzing less than half of the total Boolean space. For large circuits, this proportion is much smaller.

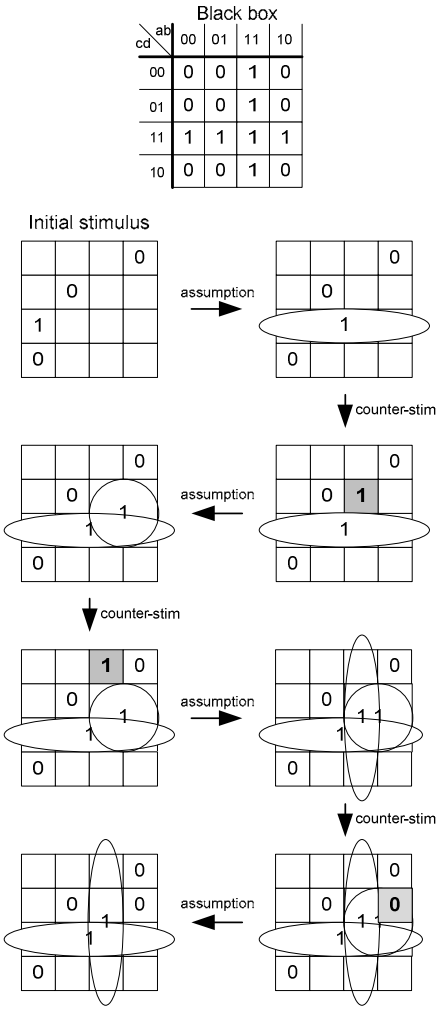


Fig. 4. Example of logic regression algorithm for function $ab + cd$

This example also demonstrates that random factor has great effect. In fact, a solution could be found as early as at the second step of the algorithm, since the ESPRESSO algorithm could implement the correct cover with the same probability.

IV. Experimental Results

For experimental studies, we used several combinational circuits from ISCAS'85 and LGSynth'89 benchmark sets. These circuits served as black boxes for the logic regression algorithm. The function of each circuit output was restored separately. The objective of the study was to show the efficiency of the developed algorithm, to see its convergence rate and applicability for medium and large circuits. Also, while running experiments, bottlenecks of the algorithm depending on the internal structure of black boxes were detected.

For all runs of the developed logic regression algorithm, the following main options were set:

- Initialization stimuli number equals 100;
- The number of stimuli on which the accuracy of the obtained solution is checked is 10000;
- The maximum number of counter-stimuli added at each iteration is 100;
- The maximum number of iterations is $K = 20$;
- The required accuracy after which the algorithm stops is $\alpha = 0.99$.

To evaluate the efficiency of the algorithm, we took into account both the final accuracy of the solution and the number of the used iterations.

Let us designate the initial (reference) circuit as R and denote its output as R_i, where i is the index of the corresponding output. As a result of the algorithm, we get a predicted circuit P, whose outputs are denoted as P_i. The accuracy is calculated separately for each output, and reflects the number of matches of the output values with the values of the corresponding outputs of the reference circuit R on an arbitrary set of input stimuli I:

$$Acc_{P_i} = \frac{|R_i(I) \leftrightarrow P_i(I)|}{|I|}, \quad (1)$$

where $|I|$ is the number of input stimuli, $R_i(I)$ is the output signature of the i-th output of the reference circuit, symbol $\leftrightarrow$ denotes vector equivalence operator. Symbols $|\dots|$ in the numerator denote the number of '1's in the vector, and therefore the number of matches with the reference.

The accuracy of the resulting circuit is defined as the arithmetic average for all outputs:

$$Acc_P = \frac{\sum_i Acc_{P_i}}{N}, \quad (2)$$

where N is the number of circuit outputs. The final success rate of logic regression for output O^i is calculated by the formula:

$$\gamma_{P_i} = Acc_{P_i} + \lceil Acc_{P_i} + 1 - \alpha \rceil \cdot \frac{I - k}{I - 1}, \quad (3)$$

where k is the actual number of iterations, I is the maximum number of iterations, α is the required accuracy, and symbols $\lceil \dots \rceil$ denote ceiling integer part. Formula (3) means that if the output for the required number of iterations does not converge, then the coefficient is equal to the achieved accuracy. And if the required accuracy is achieved, the number of iterations is also taken into account. Thus, γ_{P_i} lies within [0, 2], and is equal to 2 in the case when a complete coincidence with the reference is achieved at the first iteration. The efficiency coefficient for the whole circuit is calculated by formula (2) as the accuracy of the circuit $\gamma_P = Acc_P$.

Figure 5 shows operation process of the developed logic regression algorithm for the circuit alu4. Circuit outputs are indicated by individual curves: “o”, “p”, “q”? et. c.

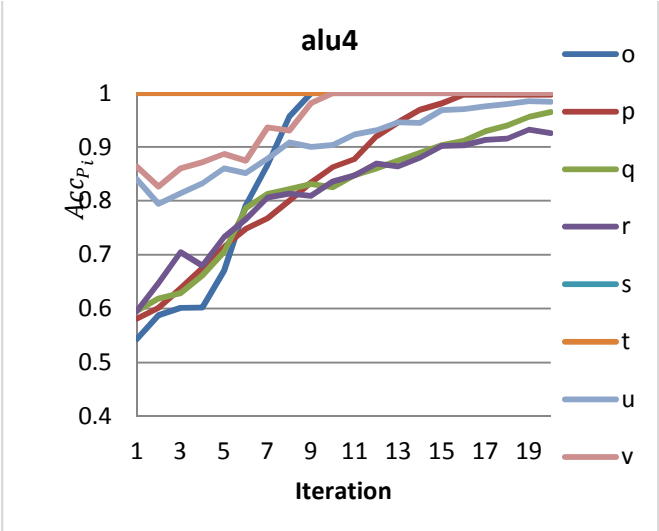


Fig. 5. Operation process of the logic regression algorithm for alu4

We see that some circuit outputs converge much faster, while others have a relatively low convergence rate and still do not reach 99% in 20 iterations. This is due to complexity of the corresponding functions. The more complex the function, the slower the algorithm converges. This complexity can be approximately estimated as the number of elements in the input cone for the given output. This is true, because all the circuits were obtained through the logic synthesis procedure, so all redundant elements have been removed. If each output is realized by the minimum possible number of elements, then we can consider this number as complexity characteristic of its function. Figure 6 shows number of elements in input cone for each output of alu4.

Relation between the number of elements in the input cone and the efficiency of the proposed algorithm is obvious. Outputs “s” и “t” were optimized at the very first iteration, while even 20 iterations were not enough for “r”, “q” и “u”.

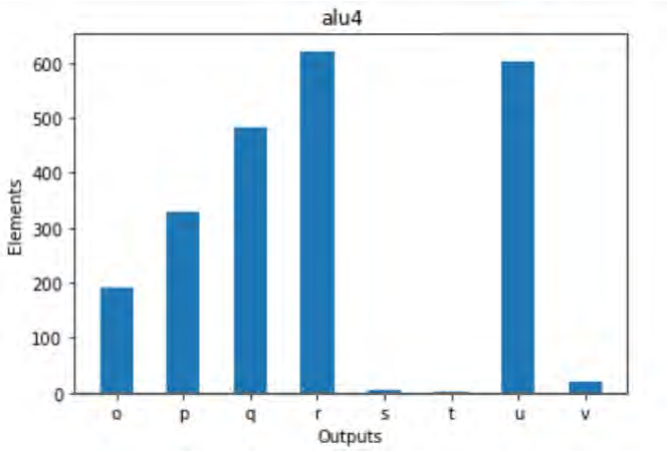


Fig. 6. Number of elements in input cones for outputs of alu4

To analyze relation between the efficiency coefficient of logic regression and the number of elements in input cone, a large number of experiments were performed on benchmark circuits from ISCAS’85 and LGSynth’89.

Dependency graph of coefficient γ_{P_i} on the number of input cone elements for different outputs for a set of arbitrary test circuits is presented in Figure 7. Pearson's correlation coefficient for these data is 0,72.

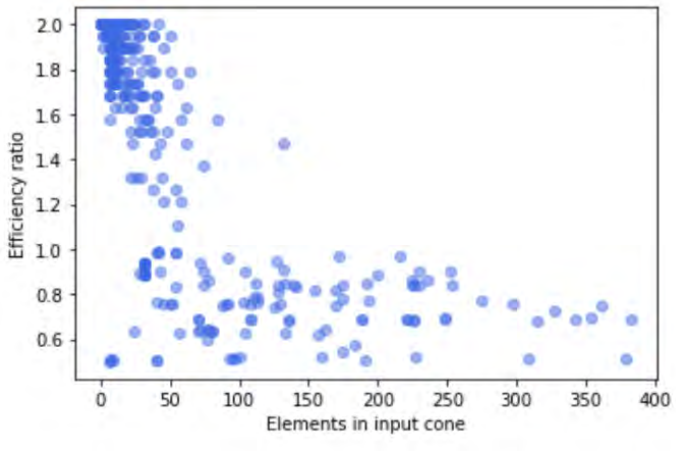


Fig. 7. Dependency of coefficient γ_{P_i} on the number of input cone elements

Dependency graph shown in Figure 7, with all the obvious correlation of the considered parameters, has some essential features. Some circuits with a small number of elements tend to have abnormally low efficiency coefficient. This is manifested by rather dense concentration of points in the lower left part of the graph. Detailed analysis of the circuits has revealed that this is largely because of the deal of XOR and NXOR elements in the output cone. In particular, the group of dots in the lower left corner of the graph refers to the outputs of the c3540 circuit, input cones of which are a tree of 7 XOR's.

V. Problems of the Current Approach and Possible Solutions

The efficiency of the proposed solution is highly limited if the circuit has many XOR and NXOR elements. The reason is that sum operations modulo two in the class of disjunctive normal forms, as well as in the class of conjunctive normal forms, cannot be minimized. That is, minimal form of DNF or CNF of XOR function of any number of variables coincides with canonic DNF or CNF, respectively. In other words, Karnaugh map for such functions contains many "checked" areas, and the correct re-conversion of such areas using this approach requires enumerating all the values inside.

One possible solution is to use approaches related to three-level minimization instead of the ESPRESSO algorithm used in this work. This will be the basis for further work in this area.

VI. Conclusion

In this paper we investigated logic regression methods for design of approximate combinational circuits. An iterative algorithm based on ESPRESSO two-level minimization was proposed. The developed software tools made it possible to recover combinational circuits with sufficient accuracy using only input/output signals. The studies showed high efficiency of the developed methods for circuits of medium size. The weak point of this approach was demonstrated, which is the complexity of two-level minimization for XOR functions. We indicate a possible solution to this problem, which is related to the use of three-level minimization methods as the core of proposed iterative method. Moreover, the use of three-level minimization will not entail restructuring of the entire flow – it is enough to replace the ESPRESSO procedure. However, other logic minimization algorithms [8] can also be used for this task - further research required.

This work was supported by the Russian Science Foundation grant No. 17-19-01645.

References

[1] Ruczinski, I. Logic Regression / I. Ruczinski, C. Kooperberg, M. LeBlanc // Journal of Computational and Graphical Statiststics. – 2003. – Vol. 12, №3 – P. 475-511.

[2] Zhang, H. Cost-Aware Patch Generation for Multi-Target Function Rectification of Engineering Change Orders / H. Zhang, J.R. Jiang // Design Automation Conference (DAC). – 2018. – P. 1-6.

[3] Gavrilov S., Ivanova G. Simultaneous Logic and Layout Synthesis for Fin-fet Based Elements with Regular Layout in Polysilicon and Diffusion // Proceedings of IEEE East-West Design & Test Symposium (EWDTS'2015), 2015, P. 264-267.

[4] Gavrilov S. V., Zheleznikov D. A., Khvatov V. M. Solving the Problems of Routing Interconnects with a Resynthesis for Reconfigurable Systems on a Chip // Russian Microelectronics, 2018, Vol. 47, No. 7, pp. 516–521.

[5] Shin, D. Approximate logic synthesis for error tolerant applications / Shin D., Gupta S. // Design, Automation & Test in Europe Conference & Exhibition (DATE 2010). – 2010. – P. 957-960.

[6] A. J. Sanchez-Clemente, L. Entrena, R. Hrbacek and L. Sekanina, "Error Mitigation Using Approximate Logic Circuits: A Comparison of Probabilistic and Evolutionary Approaches," in IEEE Transactions on Reliability, vol. 65, no. 4, pp. 1871-1883, Dec. 2016.

[7] C. Huang, C. R. Wu, T. Lee, C. J. Hsu and K. Khoo, "2019 CAD Contest: Logic Regression on High Dimensional Boolean Space," 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Westminster, CO, USA, 2019, pp. 1-6, doi: 10.1109/ICCAD45719.2019.8942137.

[8] Avdeev, N.A., Bibilo, P.N. Logical optimization efficiency in the synthesis of combinational circuits. Russ Microelectron 44, 338–354 (2015). https://doi.org/10.1134/S1063739715050029

[9] Ruczinski I., Kooperberg C., LeBlanc M. (2003) Logic Regression — Methods and Software. In: Denison D.D., Hansen M.H., Holmes C.C., Mallick B., Yu B. (eds) Nonlinear Estimation and Classification. Lecture Notes in Statistics, vol 171. Springer, New York, NY

[10] Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. Genet Epidemiol. 2005;28(2):157 - 170. doi:10.1002/gepi.20042

[11] Gavrilov S.V., Telpukhov D.V. Automated Evolutionary Design of Fault-Tolerant Logic Circuits // Problemy razrabotki perspektivnyh mikro- i nanojelektronnyh sistem (MES). 2019. № 1. P. 2-6.

[12] Miller J.F. (2011) Cartesian Genetic Programming. In: Miller J. (eds) Cartesian Genetic Programming. Natural Computing Series. Springer, Berlin, Heidelberg

[13] A. Stempkovskiy, D. Telpukhov and R. Soloviev, "Fast and accurate resource-aware functional ECO patch generation tool," 2018 Moscow Workshop on Electronic and Networking Technologies (MWENT), Moscow, 2018, pp. 1-6, doi: 10.1109/MWENT.2018.8337192.

[14] BRAYTON, R. K., HACHTEL, G. D., MCMULLEN, C. T., AND SANGIOVANNI-VINCENTELLI, A. 1984. Logic Minimization Algorithms for VLSI Synthesis. Kluwer Academic Publishers, Hingham, MA.

The Noise Immunity of CMOS Elements During their Switching and Exposure to an Ionizing Particle

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 results of modeling elements AND and OR as part of the triple majority gate are presented when switching inputs and simultaneously collecting charge from the particle track. The simulation performed using 3D TCAD physical models of CMOS transistors according to the design rule of 65 nm bulk technology with shallow trench isolation of transistor groups for tracks with linear energy transfer of 60 MeV·cm^2/mg. It was found that the beginning of switching elements AND and OR at the inputs practically does not affect the dependence of transient processes of the formation of noise pulses at the output of the element when collecting the charge from the track. The noise pulse shifted in time by a time interval equal to the time offset of the track relative to the moment of switching element inputs. Collecting the charge from the track leads to switching an element in advance of the input signals changing, either, to an additional switching delay (from -91 ps to 620 ps). At the same time, the duration of the noise pulse remains almost unchanged for the each specific track input point into the common area of the transistor location regardless of the moment of formation of the track.**

Keywords—charge collection, logical element, noise pulse, particle track, simulation, single particle

I. Introduction

CMOS combinational logic elements are the basis of encoders, decoders and majority voting logic circuits. A number of analytical papers are devoted to simulation the impacts of single ionizing particles using physics-based device models, both two-dimensional (2D) and three-dimensional (3D). In these works, it was noted [1] that the noise immunity of CMOS logic designed using bulk technology would decrease to the values of linear energy transfer by a particle on a track equal to 2 MeV·cm^2/mg when technology node will shrink to 100 nm or less. The transition of NMOS transistors to the inverse bias mode [2] and increasing the duration of the noise pulse (single-event transient) to 300–500 ps at the LET value of 30 MeV·cm^2/mg also predicted.

At technology nodes below 100 nm the CMOS logic shows the influence of diffusion transfer of charge carriers induced on the same track on adjacent circuit nodes. This joint charge collection can lead to a reduction in the duration of noise pulse [4], known in the literature as "pulse quenching". Modeling of the main characteristics of the majority voter, based on AND and OR elements designed using bulk 65-nm CMOS technology with shallow trench isolation of transistor groups was performed [5] in the our previous work.

The purpose of this work is 3D TCAD device simulation of AND and OR elements as part of a triple majority gate (TMG) designed on bulk 65-nm CMOS technology with shallow trench isolation of transistors. The purpose is to obtain quantitative estimates of the time parameters of noise pulses (single-event transients) during the simultaneous processes of elements switching and charge collection from the track of a single ionizing particle.

II. Triple Majority Gate on AND and OR Elements

Fig. 1 presents a scheme of a triple majority gate (TMG) on CMOS two-input AND logic elements (D1-D3) and a three-input OR logic element (D4). The AND (D1) and OR (D4) elements on Fig. 1 depicted as electrical schemes, but D2 and D3 AND gates as functional conventional signs. The scheme of the D1 element includes a NAND gate and an inverter. The scheme of the D4 element includes a NNOR gate and a doubled inverter.

The simulation of impacts of single nuclear particles on CMOS elements (designed on the bulk 65-nm CMOS technology) carried out using 3-D TCAD transistors models of

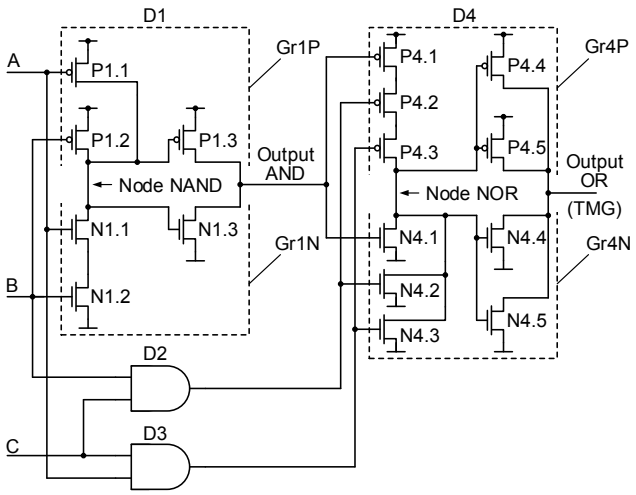


Fig. 1 Scheme of the triple majority gate based on AND gates (D1-D3) and OR gate (D4).

Funding: The reported study was funded by the Russian State assignment, project 0065-2019-0008

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

the work [6]. The 3-D device physical TCAD models of AND (D1) and OR (D4) elements presented in Fig. 2. The channel width of transistors is 400 nm for AND gates and is 800 nm for the OR gate. The designs of AND logic gates and the OR gate include groups of transistors surrounded by a shallow trench isolation with a depth of 400 nm. The shallow trench isolation covering the silicon regions of transistors to the depth of 400 nm are hidden from this picture.

AND and OR gates on Fig. 2 consist of two different groups of transistors. One group is the group of NMOS transistors Gr1N or Gr4N, another group is the group of PMOS transistors Gr1P or Gr4P. The inverter of the OR gate D4 designed as the doubled inverter (Fig. 1) with transistors pairs (N4.4, N4.5, and P4.4, P4.5) located upon opposite sides of the NOR gate in groups Gr4N and Gr4P (Fig. 2). The results of simulation obtained by using Sentaurus Device at the temperature 25°C and the supply voltage of 1.0 V for particle tracks with linear transfer energy 60 MeV·cm^2/mg.

III. Switching AND and OR Elements ahead of time of Input Signals Changing

Fig. 3a shows voltages on nodes of the AND element during the time charge collection from the track T1N passing through NMOS transistors of the group Gr1N. The signals at inputs of the triple majority gate switch from A = B = C = 0 to A = B = 1, C = 0. Linear energy transfer to the track is LET = 60 MeV·cm^2/mg, the start of charge collection at t_{TR} =160 ps. In the case of track T1N passing through closed NMOS transistors of the NAND element the transistors go to the inverse bias mode. Charge collection by them from the track leads to switching the element (Output AND on Fig. 3a) ahead of time of changing of input signals.

The NMOS transistor of the inverter closes and begins to collect the charge from the track T1N, forming after advance switching a noise pulse of negative polarity at the AND output with an amplitude of 0.7 V (Fig. 3a). It should be noted that advance switching of the AND gate leads to advance switching of the OR gate (curve "Output OR" in Fig. 3a) before time changing the input signals A = B of TMG on 31 ps less. As a result we have the negative value of the delay time of the AND element t_{DL} = -31 ps.

In the case of the OR element (Fig. 3b), when the input signals of the triple majority gate switched from A = B = 1, C = 0 to A = B = C = 0, only one closed PMOS transistor P4.3 of the node NOR collects a charge from the T4P track passing through the group Gr4P (Fig. 2).

The collected charge is enough to charge the NOR node to the voltage of 0.8 V (Fig. 3b) and during this time the inverter switches of the output OR to logical level "0" before changing the input signals A = B of the triple majority gate. After beginning of a change the input signals A = B (1→0) the NOR node starts to charge to the level of 1 V through the open PMOS transistors of the NOR group. As a result we have the negative value of the delay time of the OR element t_{DL} = -30 ps. The of switching times from the logical level "0" to the logical level "1" (the example is in Fig. 3a) or from the logical level "1" to the logical level "0" (the example is in Fig. 3b) are from 9 ps to 12 ps.

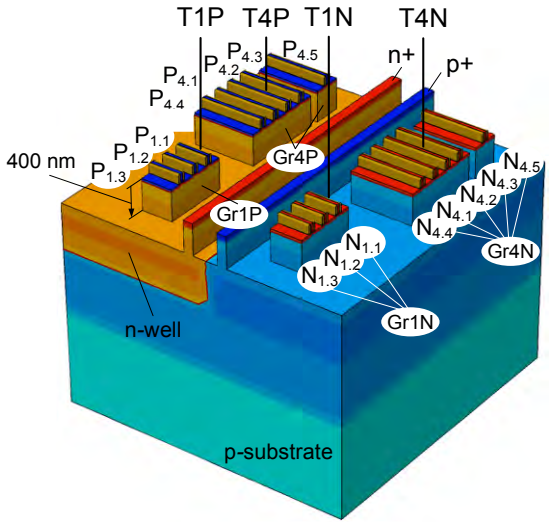


Fig. 2. 3-D device physical TCAD model of the AND (D1) and OR(D4)-elements of the triple majority gate; directions of tracks T1P, T1N, T4P and T4N are at the normal to chip surface; n+ and p+ regions are the parts of the guard rings.

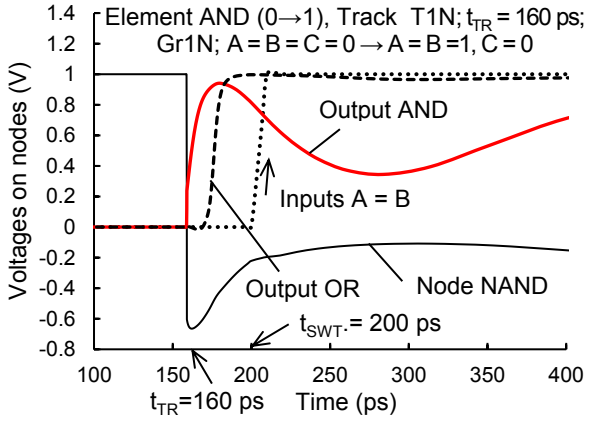


Fig. 3. Voltages on nodes for the case of advanced switching of elements ahead of input signals changing, tracks with LET = 60 MeV·cm^2/mg, the formation of the track at t_{TR} = 160 ps, switching inputs at t_{SWT} = 200 ps: (a) AND gate, the input track point is T1N at the group Gr1N, switching from A = B = C = 0 to A = B = 1, C = 0; (b) OR, the input track point is T4P at the group Gr4P, switching from A = B = 1, C = 0 to A = B = C = 0.

IV. Switching of Elements with Additional Delay

The curvers on Fig. 4a for the AND element characterize additional switching delays when the charge is collected by the PMOS transistors of the NAND group. In the case of the track T1P passing through the PMOS transistors of the NAND group PMOS transistors initially remain open until the input signals of the TMG are switched from “0” to “1”.

After switching input signals of the TMG from “0” to “1” the voltage of 1 V is set on gates of PMOS transistors P1.1, P1.2 of the NAND group (scheme of the AND element on Fig. 1). The voltage of 1 V is stored on the drains of transistors P1.1, P1.2 and the gate of the transistor P1.3 of the inverter (Fig. 4a).

Then the NAND node begins slowly to discharge by the current of the series-connected NMOS transistors of the NAND group, and this change is inverted to the output by the inverter of the AND element until it and OR output (TMG output) are set to 1 V. As a result we have the delay time of the switching of the AND element t_{DL} = 286 ps.

It should be noted that when the AND element switches with an additional delay, the OR element of this TMG has the less the delay time to switch (the curver “Output OR” on Fig. 4a).

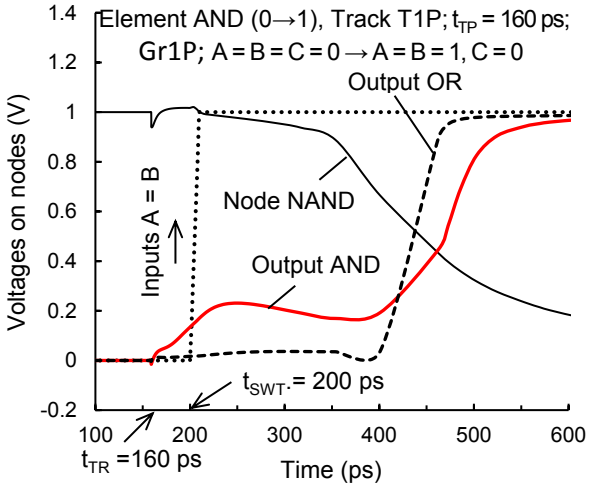
The curvers on Fig. 4b for the OR element characterize additional switching delays when the charge is collected by the NMOS transistors of the NOR group. NMOS transistors of the NOR group (Fig. 1) are in open state before switching the OR element. When collecting the charge from the track T4N, NMOS transistors of the NOR group go to inverse bias mode, and NMOS transistors of the inverter remain closed, collecting the charge of the electrons. This decreases the voltage at the OR output (Fig. 4b). Only after the NMOS transistors exit from the inverse bias mode, the output of the OR element returns to the logical zero level “0”. As a result we have the delay time of the switching of the OR element t_{DL} = 566 ps.

V. Switching Delays of AND and OR Elements relative to Moment of TMG Input Signals Changing

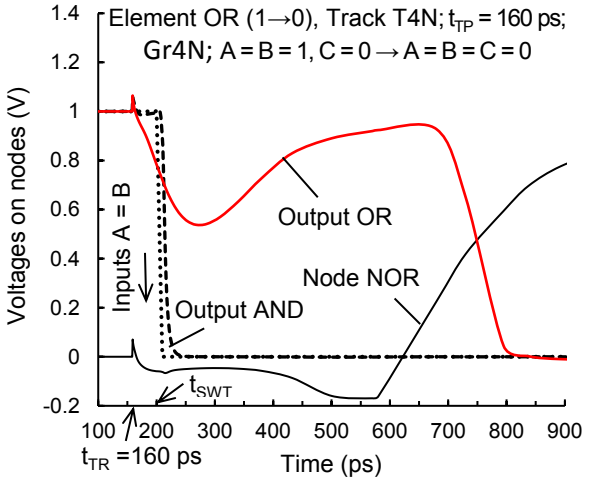
Fig. 5 shows switching parameters of AND and OR elements as functions on the time of formation of the track of a single ionizing particle with LET = 60 MeV·cm^2/mg. The curves have notations corresponding to AND or OR gates, to the nature of switching is 0→1 or 1→0, the types of track are T1N, T1P, T4N, and T4P.

These parameters form two groups. The first group (Fig. 5a) characterized switching of the logical elements of the TMG in ahead of the time before the signals at TMG inputs change. Advance switching is caused by collecting the charge from the track by initially closed transistors of the NAND node (and NOR node). Delays for cases of advance switching have negative values of the time before the signals at TMG inputs change in t_{SWT} = 200 ps (Fig. 5a).

Advance switching in time occurs almost identically when collecting charge from tracks through NMOS and RMOS transistors of AND (OR) elements, as well as when switching the inputs of the majority gate from the state "0" to "1" and from "1" to "0" (Fig. 5a). The of switching times from the output logical level "0" to "1" or from "1" to "0" are from 9 ps to 12 ps.



(a)



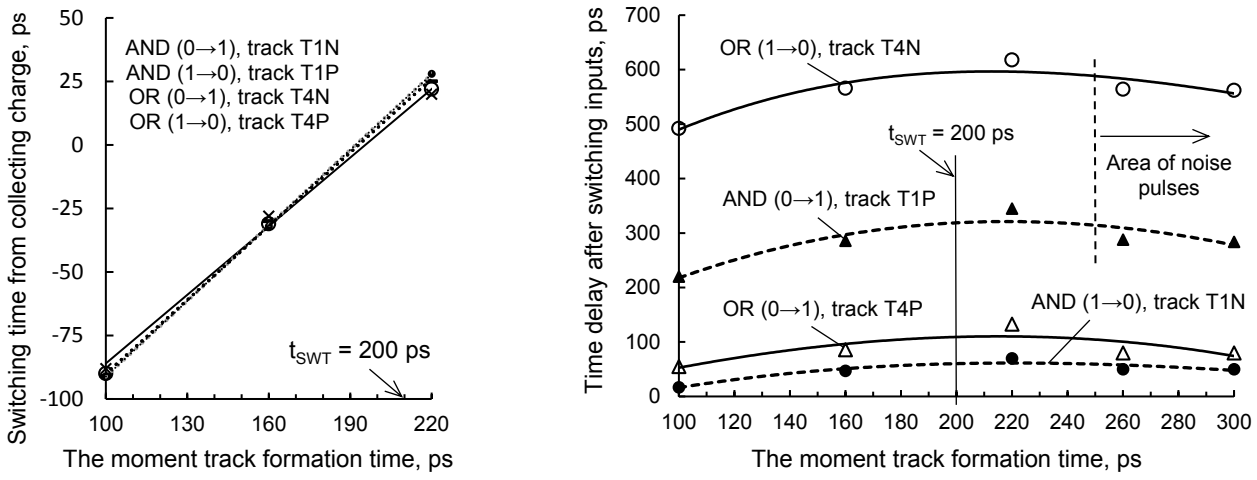
(b)

Fig. 4. Voltages on nodes of elenents forming the TMG for case with an additional switching delay when collecting charge from the track with LET = 60 MeV·cm^2/mg, the formation of the track at t_{TR} = 160 ps, switching inputs at t_{SWT} = 200 ps: (a) AND gate, the input track point is T1P at the group Gr1P, switching from A = B = C = 0 to A = B = 1, C = 0; (б) OR gate, the input track point is T4N at the group Gr4N, switching from A = B = 1, C = 0 to A = B = C = 0.

Therefore, the time of advanced switching is practically the same for a specific time of the track formation (Fig. 5a).

The second group (Fig. 5b) characterized switching of the logical elements with an additional delay caused by the charge collection from the track by the initially open transistors of the NAND node (NOR node) that closed after the signals at TMG inputs change in t_{SWT} = 200 ps.

In both cases, when the state changes, at the initial moment of charge collection the transistors of the NAND (NOR) node go into inverse bias mode. This switches the inverter of the AND (OR) element so that the closed transistor of the inverter after such a change in the state begins to collect charge from the track, forming a noise pulse (voltage drawdown) at the AND (OR) output.



(a) (b)

Fig. 5. Switching characteristics of AND and OR elements depending on the time of formation of the track of a single ionizing particle with LET = 60 MeV·cm^2/mg: (a) switching time from collecting charge in advance of the input signals changing; (b) switching after changing inputs with an additional time delay. Characteristics have symbols corresponding to the element AND or OR, the type of input switching 0→1 or 1→0 and the type of track T1N, T1P, T4N и T4P.

The switching delays for cases of advance switching have negative values (Fig. 5a) and decrease when the time of track occurrence approaches to the moment when the signals at TMG inputs change.

For cases when input signals of TMG change at t_{SWT} = 200 ps and the track is formed at t_{TR} = 220 ps, the switching delay values are determined by own delays of AND (OR) elements without charge collection. For tracks with t_{TR} = 260 ps and 300 ps, after switching TMG inputs there is no delay, but the formation of a noise pulse, in these cases the TMG input signals remain unchanged. In general, switching delays can range from negative till positive values -91 ps to 620 ps.

VI. Simulation Results

The main results of the modeling:

Durations of noise pulses at the outputs of the AND and OR elements of the triple majority gate, when switching TMG input signals and at the same time collecting the charge from the particle track, practically do not depend on the moment of track formation for specific input track points and signals at the TMG inputs. This is typical for tracks directed to the region of a group of transistors with the same conductivity type, located in a common silicon area, bounded by a shallow trench isolation for a specific input track point.

Transistors of NAND (NOR) groups, which are in the closed state at the specified input signals of AND (OR) elements, switch when collecting the charge from the track, leading to switching of the TMG in ahead time before changing the signals at its inputs.

The transistors of NAND (NOR) groups, which are open by the input signals of AND (OR) elements, do not switch when the charge is collected by them from the track. At the same time, their charge collection essential delays switching of the TMG despite changing its input signals.

Collecting the charge from the track of a single nuclear particle formed after the completed switching of the majority gate leads to the formation of a noise pulse at the output of the element AND (OR) and at the output of the majority gate. The noise pulse durations are maximum for tracks that pass through initially open transistors with common areas of their drains in the PMOS transistors group of AND element and in the NMOS transistors group of OR element. Delays in these cases characterized by values of the logical elements with an additional delay during switching with a charge collecting.

VII. Conclusion

The presented features of the elements is useful when designing CMOS microprocessor systems for space applications. In particular, this is an advance switching of the triple majority gate or an additional increase in its switching delay, initiated by the charge collection from the track, depending on the state of TMG inputs. The switching delay can vary from -91 ps to 620 ps depending on the input track points to the chip and the input signals of TMG.

References

[1] P.E. Dodd, M.R. Shaneyfelt, J.A. Felix, and J.R. Shwank, "Production and propagation of single-event transients in high-speed digital logic ICs", *IEEE Transactions on Nuclear Science*, 2004, vol. 51, no. 6, pp. 3278–3284.

[2] P.E. Dodd, and L.W. Messengill, "Basic mechanisms and modeling of single-event upset in digital microelectronics", *IEEE Transactions on Nuclear Science*, 2003, vol. 50, no. 3, pp. 583–602.

[3] V. Ferlet-Cavrois, L.W. Messengill, and P. Couker, "Single-event transients in digital CMOS – A review", *IEEE Transactions on Nuclear Science*, 2013, vol. 60, no. 3, pp. 1767–1790.

[4] N.M. Atkinson, A.F. Witulski, W.T. Holman, J.R. Ahlbin, B.L. Bhuva, and L.W. Massengill, "Layout technique for single-event transient mitigation via pulse quenching", *IEEE Transactions on Nuclear Science*, 2011, vol. 58, no. 3. pp. 885–890.

[5] Yu.V. Katunin, and V.Ya. Stenin, "Modeling of single ionizing particles impact on logic elements of a CMOS triple majority gate", *Russian Microelectronics*, 2020, vol. 49, no. 3, pp. 214–223.

[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.

Noise Reduction in Reset Domain Crossings Verification Using Formal Verification

Mohamed Fawzy, Ahmed Elgohary, Hala Ibrahim

Mentor, A Siemens Business, Cairo, Egypt

mohamed_fawzy@mentor.com, ahmed_elgohary@mentor.com, hala_ibrahim@mentor.com

***Abstract*— Reset architecture of a digital design can be quite complex. Typically, SoC designs have multiple sources of reset, such as power-on reset, hardware resets, debug resets, software resets, and watchdog timer reset. These multiple reset domains make the design potentially exposed to metastability issues, so the designer must perform the reset domain crossing (RDC) analysis and resolve any RDC issues in the early stages of designing. This can quite be challenging, because of the effort needed for this analysis and how noisy it can be. In this paper, we present some of the challenges in the existing methodology for RDC analysis and propose a new methodology to reduce RDC results noisiness and achieve more accurate results. This leads to faster verification closure. The results are concluded by applying the proposed methodology on a set of real designs.**

Keywords*— *Formal Verification, Metastability, RDC, Reset Domain, Reset Verification

I. Introduction

With the increased complexity of digital designs, designs reset architecture has also become very sophisticated. While implementing such complex architecture, designers tend to make some mistakes, which can lead to metastability, glitches, or other functional failures in the system. Reset domain crossing (RDC) refers to a sequential path in the design where the source and the destination sequential elements operate on different independent resets. Metastability happens when an asynchronous reset from one reset domain causes a transition too close to the clock edge of a flip-flop in another reset domain or without a reset causes a non-deterministic flip-flop value that propagates throughout the design resulting in functional failures. For example, having a signal traveling between two registers, each has a different asynchronous reset domain as shown in Figure 1. In this case, the asynchronous reset of the transmitter register can change the register output within the metastability window of the receiving register.

Currently, various tools in the market perform RDC analysis. In this paper, results are obtained and investigated using Questa-RDC. The purpose of the RDC analysis is to resolve the issues associated with signals propagating from one reset domain to another, which are considered potential sources of metastability. Solutions of RDC issues could be synchronization, isolation, or by using reset ordering. Reset ordering means that the metastability issue can be avoided if the destination flop can be held in the reset state before asserting the source reset signal. There is no probability for RDC issue if both source and destination registers are operating in the reset state simultaneously even if they belong to different reset domains. Reset ordering constraints are known user constraints that list the ordered resets and given to the RDC analysis tools to ignore the crossings between the ordered resets. There is a different category of reset signals which is the logic dependent resets. In this category, both source and destination registers are operating in the reset state simultaneously as the reset combinational logic of destination register will always assert once the reset of the source register is asserted. RDC analysis might be challenging because of the following reasons:

1) Results can be noisy due to not considering dependent resets that do not cause metastability issues.
2) Significant time and effort needed for the detection of such dependent resets as the reset assertion logic reaching up to source and destination registers can be complex combinational logic.
3) Verification effort needed to verify that these resets are completely dependent in all cases.
4) The manual effort needed to set up the design including information of dependent resets to impact RDC analysis.

Formal verification is the primary term for a group of techniques that use static analysis based on mathematical transformations to determine the correctness of hardware or software behavior in contrast to dynamic verification techniques such as simulation. As design sizes have increased and accordingly simulation times, verification teams seek for ways to decrease the number of test cases needed to exercise the system to have an acceptable coverage degree.

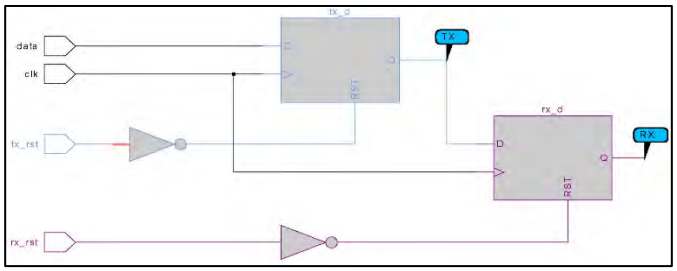


Figure 1 Reset domain crossing

Formal verification tools use various algorithms to verify the design and do not perform any timing checks. These tools do not require a stimulus or a testbench, and thus, formal verification is performed early in the IC design cycle as soon as the RTL code is available. The sooner a bug is found, the easier it is to fix. One of the most beneficial incomes of using formal verification is it is considered an exhaustive methodology that covers all input scenarios and also detects corner-case bugs.

No previous work has addressed the challenge of noise reduction of RDC results due to having dependent resets. These dependent reset related issues should be detected and optimized such that they should not be considered RDC violations. On the other hand, the work in [1] addressed the verification of RDC results using assertion-based formal verification technology.

The paper is organized as follows. Section II explains the proposed methodology of noise reduction of RDC results due to reset dependencies. Section III shows applying the methodology on different examples of reset domain dependencies. Section IV explains the different categories of reset dependencies. Section V shows the results of our proposed methodology on real case studies. Finally, section VI concludes the paper.

II. Proposed Methodology

The proposed methodology is shown in Figure 2 going through the following steps:

1. RDC Analysis. This is performed using Questa-RDC to extract the list of the reset domain crossings in the design.
2. Automated extraction of source and destination reset pairs of all crossings. This is done using a script, which parses the design and the results of RDC analysis. In Figure 3, source reset is extracted as “r0” and destination reset is extracted as “r2”. These reset pairs will be parsed and passed to another script, which generates the required assertions.
3. Automated assertions generation to check the dependency of Tx/Rx Resets. Known assertion statements are automatically generated using the extracted reset pairs from the previous step.
4. Automated assumptions generation of user constraints for known ordered resets. These assumptions are to be considered during formal verification analysis.
5. Automated formal setup generation. This step to automate the generation of Formal setup files to be used during formal verification analysis.
6. Formal Verification. This is through running formal analysis to identify proven and fired assertions. Proven assertions mean that the reset pair is completely dependent. Accordingly, the related RDC crossing is safe and should be filtered out from RDC verification.

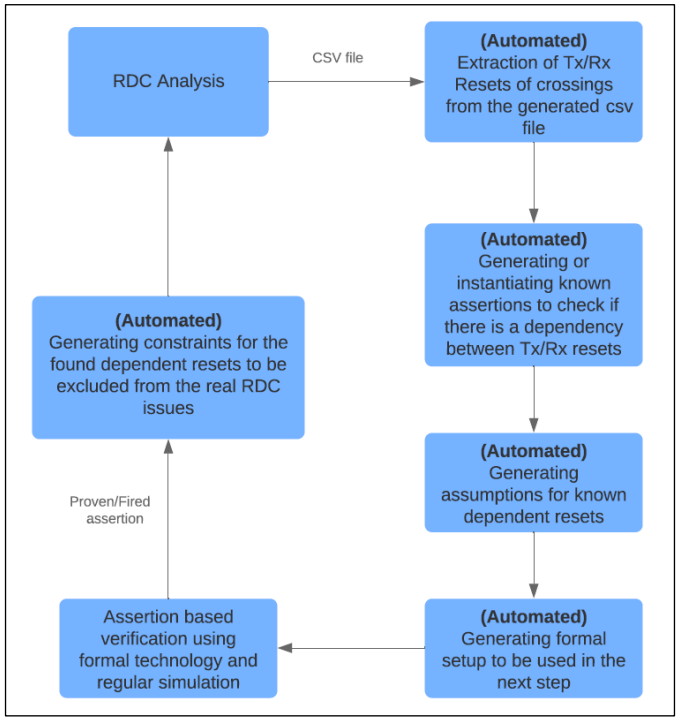


Figure 2 Proposed methodology flow

Firing assertions mean that the reset pair is independent and there’s a probability for metastability. Accordingly, RDC crossing should still be reported in RDC verification.

7. Automated constraints generation for the found dependent resets. This is done by parsing proven assertions from the formal analysis and generating of reset ordering constraints for the reset pairs to exclude related crossings between such dependent resets in the next RDC analysis.
8. Rerunning RDC analysis after considering the generated constraints from the previous step. Results should be more accurate and less noisy after filtering the dependent resets crossings.

III. Applying the Proposed Methodology on Different Scenarios

A. *RDC Crossing with Tx/Rx Dependent Resets*

The RDC shown in Figure 3 has dependent Tx/Rx Resets. Once Tx reset ”r0” is asserted “active low”, Rx reset “r2” will also be asserted “active low”. The generated assertion is shown in Figure 4.

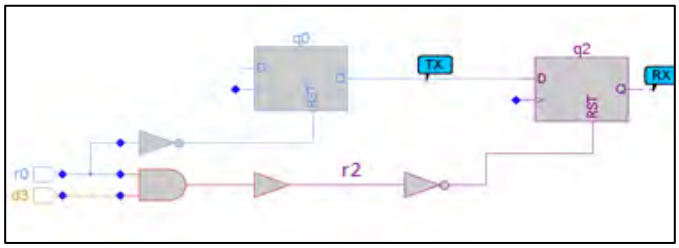


Figure 3 Dependent resets example

```
property prop_rdc_ordered (tx_rst, rx_rst, clk);
  @(posedge clk) disable iff (`CONFIG_MODE)
      tx_rst |-> rx_rst;
endproperty
assert property (prop_rdc_ordered (~r0,~r2,clk));
```

Figure 4 Assertion used to verify dependent resets

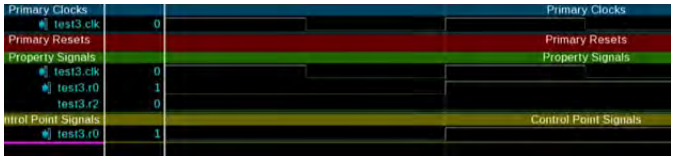


Figure 5 Formal verification results for dependent resets

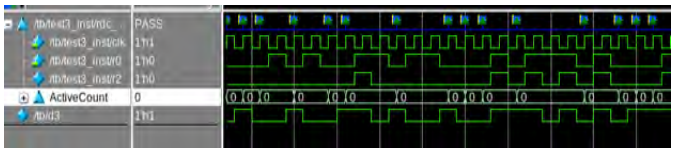


Figure 6 Simulation results for dependent resets

Formal verification proves this assertion as shown in Figure 5 and the simulation result for the same assertion is shown in Figure 6. This means these resets are dependent resets and such RDC crossing will not cause metastability issues. Accordingly, formal design constraints are generated and applied to the tool set up to avoid reporting this false crossing.

B. RDC Crossing with Tx/Rx Independent Resets

The RDC in Figure 7 has independent Tx/Rx Resets. Once Tx reset "r0" is asserted "active low", Rx reset "r2" will be de-asserted "active high". The generated assertion is shown in Figure 8.

Formal verification fires this assertion as in Figure 9, which means these resets are independent resets and such RDC crossing will cause metastability issues and accordingly should be reported as a real RDC violation. The simulation also fires with any given stimulus and this is shown in Figure 10.

C. RDC Crossing with Known ordered Resets

The RDC in Figure 11 between Tx/Rx resets have different signals in their fan-in. Tx reset is a combinational logic of "rst1" and "rst2". Rx reset is a combinational logic of "rst2" and "rst3". Once "rst1" is asserted, both Tx and Rx resets are asserted. Metastability happens when "rst2" is not asserted and "rst1" is asserted while "rst3" is not asserted. The generated assertion for this case is shown in Figure 12.

Formal verification fires this assertion as shown in Figure 13, which means these resets are independent resets and such RDC crossing will cause metastability issues. The simulation result shows an assertion firing as well in Figure 14.

There is a user constraint that "rst1" and "rst3" are ordered resets. This means that "rst1" and "rst3" are asserted simultaneously. Our proposed methodology automatically generates the assumption shown in Figure 15 based on the given user constraint. After applying this assumption, the assertion correctly proves to indicate that Tx/Rx resets are dependent resets.

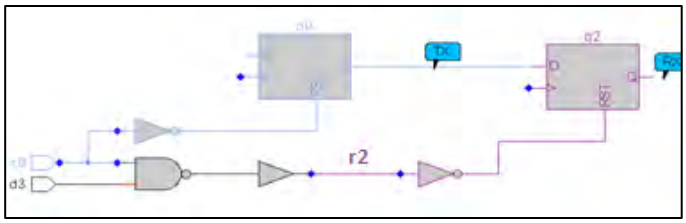


Figure 7 Independent resets example

```
property prop_rdc_ordered (tx_rst, rx_rst, clk);
  @(posedge clk) disable iff (`CONFIG_MODE)
      tx_rst |-> rx_rst;
endproperty
assert property (prop_rdc_ordered (~r0,~r2,clk));
```

Figure 8 Assertion used to verify independent resets

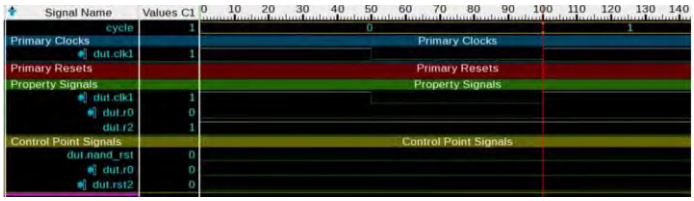


Figure 9 Formal verification results for independent resets

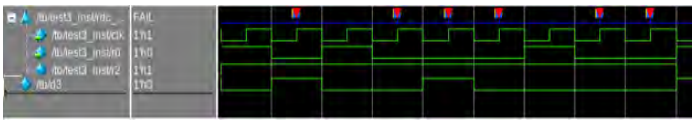


Figure 10 Simulation example of assertion firing for independent resets

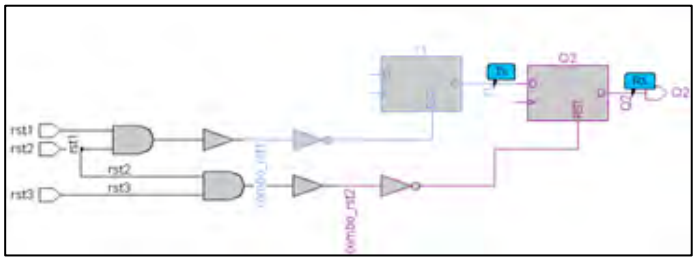


Figure 11 Known dependent resets example

```
property prop_rdc_ordered (tx_rst, rx_rst, clk);
  @(posedge clk) disable iff (`CONFIG_MODE)
      tx_rst |-> rx_rst;
endproperty
assert property (prop_rdc_ordered (!combo_rst1, !combo_rst2, clk));
```

Figure 12 Assertion used to verify RDC with known dependent resets

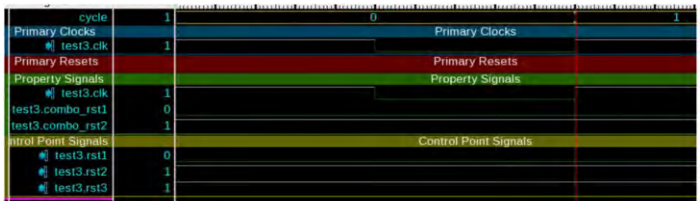


Figure 13 Formal verification results for known dependent resets

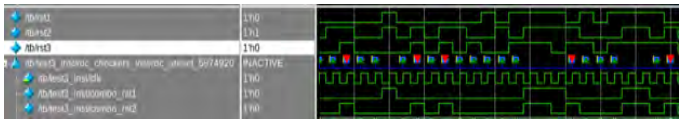


Figure 14 Simulation example for assertion firings of known dependent resets

```
property assume_rdc_ordered (tx_rst, rx_rst, clk);
  @(posedge clk)
     tx_rst = rx_rst;
endproperty
assume property (assume_rdc_ordered (!rst1, !rst2, clk));
```

Figure 15 Generated assumption based on user constraints

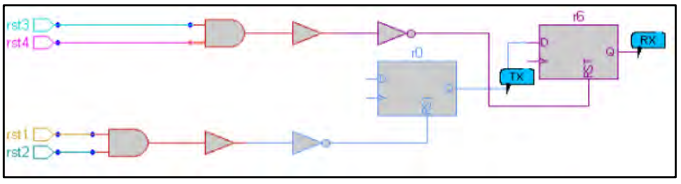


Figure 16 No common Signals between Tx/Rx resets

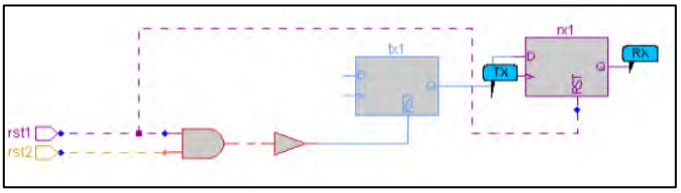


Figure 17 Rx reset is a subset of Tx reset supports

IV. CATEGORIES OF RESET DEPENDENCIES

A. All the supports of Tx reset are a subset of Rx reset supports

Figure 3 shows an example of this category, where Tx reset is a subset of fan-in of Rx reset. The Dependency of resets is depending on the combinational logic and polarity of resets.

B. Tx reset and Rx reset have some common supports but also extra supports

Figure 11 shows an example of this category, where Tx/Rx reset have common supports “rst2” and also extra supports “rst1”, “rst3”. The Dependency of resets depends on reset ordering constraints of extra supports.

C. No common signals between Tx/Rx resets

Figure 16 shows an example of this category, where no common signals between Tx and Rx resets. The Dependency of resets depends on reset ordering constraints of extra supports.

D. Rx reset are a subset of Tx reset supports

Figure 17 shows an example of this category, where Rx reset is a subset of fan-in of Tx reset. The Dependency of resets depends on the combinational logic and polarity of resets.

V. CASE STUDIES AND RESULTS

In this section, we are going to demonstrate how our proposed methodology affects the reduction of false RDC results in real designs after applying correct design constraints, assumptions, and waiving non-real RDC violations.

A. Case Studies.

i) Design A

This design has 41 RDCs, thus 41 assertions have been generated and tested using Formal Verification. After running Formal Verification on these assertions, 14 assertions are fired (independent resets) and 27 assertions are proven (dependent resets).

Figure 18 shows an example; once Tx reset “rst1” is asserted, Rx register will be asserted. Thus we can say that Tx and Rx resets are dependent.

ii) Design B

This design has 158 RDCs, thus 158 assertions have been generated and tested using Formal Verification. After running Formal Verification on those assertions, 147 assertions fire (independent resets) and 11 assertions are proven (dependent resets).

In Figure 19, Tx reset is OR-ed with another signal to create the Rx reset, thus if the Tx register is asserted, the reset signal of the Rx register will be asserted as well and there will be no chance of metastability.

In Figure 20, if Tx reset “rst1" asserts the output of “r1” register will go high, which will pass from the OR gates to assert the Rx register reset signal. Thus, if the Tx register is asserted, the reset signal of the Rx register will be asserted and there will be no chance of metastability

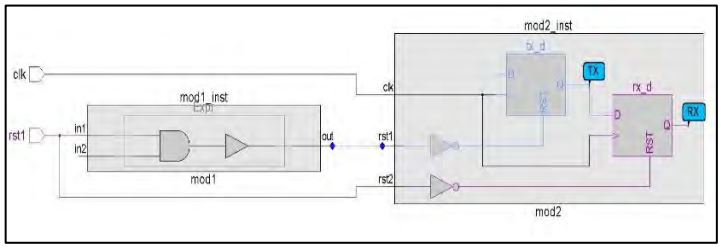


Figure 18 Design 1 - RDC 1 (Simplified schematic)

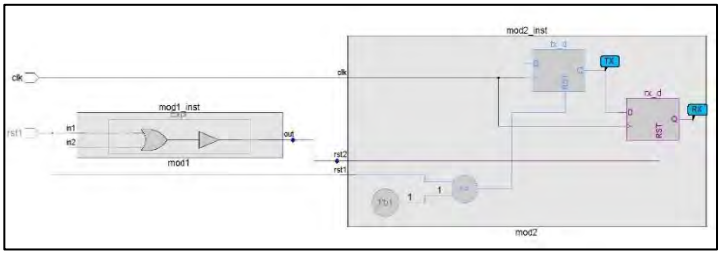


Figure 19 Design 2 - RDC 1 (Simplified schematic)

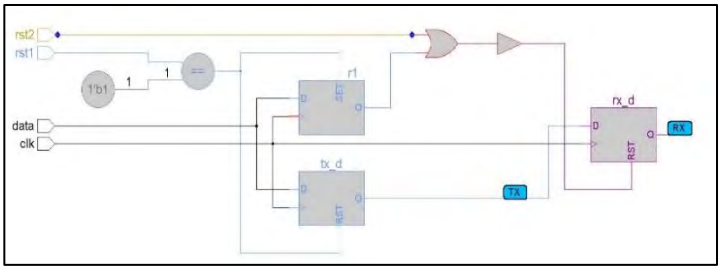


Figure 20 Design 3 - RDC 2 (Simplified schematic)

B. Results

Numbers in Table 1 shows the number of RDC violations before and after running formal verification and differentiating between correct and false (noisy) RDC issues, hence, applying the correct issues waivers or design constraints required.

TABLE I
NUMBERS OF RDC VIOLATIONS BEFORE AND AFTER APPLYING THE PROPOSED TECHNIQUE

Design Number	#total rdc_areset/rdc_cdc_areset crossings	#filtered rdc_areset/rdc_cdc_areset crossings	Filtration percentage
1	67534	5373	8%
2	24603	3937	16%
3	5878	675	11%
4	21800	126	0.60%
5	782	60	8%
6	111	28	25%
7	41	27	71%
8	1140	16	1.50%
9	19	12	6%
10	158	11	7%
11	158	11	7%
12	13	11	85%
13	9	9	100%
14	9	9	100%
15	80283	9	0.10%
16	188	2	1%
17	2	2	100%
18	7	1	14%

VI. CONCLUSION

The proposed flow showed that we can reduce the number of the reset domain crossings by excluding the noisy false crossings, which propagate through two dependent reset domains, thus reducing the noisiness of the results and the needed time for analysis. This is done through the automatic generation of design constraints/assumptions and/or reset ordering constraints after running formal verification that differentiates between the correct and noisy RDC issues.

VII. REFERENCES

[1] Ahmed, K. Nouh, and A. Abbas, "Multiple reset domains verification using assertion based verification," 2017 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), Abu Dhabi, 2017, pp. 1-6.
[2] Questa® RDC User Guide, version 2020.2, 2020.
[3] Questa® PropCheck User Guide, version 2020.2, 2020.

Typical Signal Correction Structures Based on Duplication with the Integrated Control Circuit

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

Dmitry Efanov,
DSc, Professor at Higher School of Transport,Institute of Mechanical Engineering, Materials and Transport, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation
TrES-4b@yandex.ru

Abstract—The paper describes the research results in the field of developing the methods of the synthesis of fault-tolerant discrete devices. The authors propose to use the signal correction structure based on duplication with the integrated control circuit in the synthesis of discrete devices. A generalized structure of signal correction based on duplication and control by some diagnostic feature is described. Three typical signal correction structures based on duplication are introduced. The first structure is based on the control of calculations by repetition codes. The second structure involves the control of calculations by parity codes. The third structure is based on the control of calculations by a special code with summation of weighted transitions, which was also developed by the authors. In the first structure, any errors are detected at the outputs of the controlled object. In the second structure, any errors with odd multiplicities are detected. In the third structure, any errors are detected, except for errors with the maximum multiplicity (which distort all signals). In the experiment with MCNC Benchmarks, it is shown that the first structure is comparable in implementation complexity to the traditional structure based on triple modular redundancy. Moreover, the structure with the control by repetition codes is slightly inferior to it in complexity. The second structure is much simpler, but correction of any errors is not guaranteed. The third structure makes it possible to construct simpler fault-tolerant devices, while guaranteeing the detection of any errors, with the exception of errors associated with distortions of all output signals of the checking device. Such errors are very rare on the outputs of real devices. The second structure can be used for checking groups of independent outputs by parity codes in the control circuit, which allows detecting any errors in the controlled object. The proposed duplication-based signal correction structures are constructed from standard blocks, which allow them to be used widely enough for the synthesis of fault-tolerant discrete devices.

***Keywords*—fault-tolerant discrete devices; signal correction circuit; majority signal correction structure; integrated control circuit; control of calculations by parity codes; control of calculations by a code with summation of weighted transitions; structural redundancy of the device.**

I. INTRODUCTION

The methods of faults detection and correction of incorrectly calculated values (signals) are widely used in the construction of reliable and safe discrete systems [1 – 4]. These methods are often used comprehensively. The structures of discrete systems with the properties of faults detection and correction of incorrect signals are based on the well-known principles of noise-resistant and error-tolerant coding [5, 6]. For example, they are provided with self-checking integrated control circuits for the synthesis of systems with fault detection. Block separable and non-separable codes focused on the detection of distortions are used to organize this [7]. Such codes, for example, include various sum codes [8] and constant-weight codes [9]. The codes focused on error correction in the bits of code words are used in the synthesis of systems with correction of calculation results [10, 11]. It is known from the coding theory that it is necessary to provide a Hamming distance $d \geq 2d_c + 1$, where d_c is the multiplicity of the corrected error, to correct errors with a multiplicity of d. For example, if d_c=1, the Hamming distance should be d≥3. A similar principle is used in typical signal correction structures. The commonly used structure of the majoritarian signal correction is based on triplication of source blocks with checking the calculations of the same-name outputs on the majority elements [12]. This ensures that the device is not sensitive to single fault manifestations.

The structure of signal correction based on triplication with the majority principle of choosing the correct values is used in all branches of science and technology. It is used in the development of highly reliable control systems for responsible technological processes both in the industrial and transport sectors [13 – 19].

The structure with the majority principle of choosing the correct values, despite its advantages associated with the ability to correct the manifestations of errors in calculations, has a significant disadvantage. It is associated with the significant introduced redundancy: three source blocks instead of one, as well as a signal correction circuit are required. In addition, often the source block and its copies are provided with a self-checking integrated control circuits to identify the incorrectly functioning blocks, and the correction circuit itself is synthesized in the form of a self-checking device.

In this paper, the authors highlight the results of research on the possibilities of synthesizing the signal correction circuits with reduced complexity of technical implementation in comparison with the traditional structure with triple modular redundancy. The article proposes new typical signal correction structures based on the duplication and the application of the integrated control circuits by binary redundant codes.

978-1-7281-9899-6/20/$31.00 ©2020

II. THE MAJORITY SIGNAL CORRECTION STRUCTURE

The classical structure of signal correction based on triple modular redundancy is shown in the Fig. 1. To achieve the fault tolerance property regarding to single faults, this structure uses the main device $F(x)$ and two copies of it – $F^*(x)$ и $F^{**}(x)$. All three devices work in parallel and implement the same functions on the same input influences. The values of signals from the same-name outputs are compared at the inputs of the majority elements, forming a signal correction circuit. These elements are insensitive to distortion at the inputs, as well as to their own faults before the output cascade. With this feature of majority elements in mind, highly reliable components are often used for their implementation. In addition, methods for the synthesis of self-checking majority elements are known [20]. Let's name a structure with triple modular redundancy the *M*-structure.

Triple modular redundancy in the *M*-structure is necessary to give it the property of insensitivity to the manifestations of faults in the source device in the form of signal distortions at its outputs. Its use in practice leads to a significant increase in the complexity of the technical implementation of the final device. In addition, the above structure has a drawback associated with the inability to identify an incorrectly functioning device. This problem is solved by the retrofitting of each unit of the self-checking integrated control circuits, which, however, leads to the complexity of the circuit as a whole.

It should be noted that equipment diversification is used as an additional means of increasing the fault tolerance and the ability to fix a wide class of faults (components that perform identical functions are implemented according to different principles, or the modes and algorithms of the system functioning change, execution time reserves for operations are introduced, etc. [3, 21]).

Several cases may occur when developing a structure with the majority principle of correction. The first case is when the original device $F(x)$ is given to the developer in unchanged form, and its task is to develop a fault-tolerant system, while the developer can use exact copies of this device. The second case is when it is possible to optimize two additional copies of the $F(x)$ device during the development of a fault-tolerant system. The third case is when it is possible to optimize the $F(x)$ device itself and its copies. In this case, it is obvious that of the three options, the last one will provide the least structural redundancy. Thus, for typical fault-tolerant structures, it makes sense to introduce the concept of *a structure with the minimal redundancy*. This is a structure that will provide the least technical implementation complexity for the selected option of introducing redundancy.

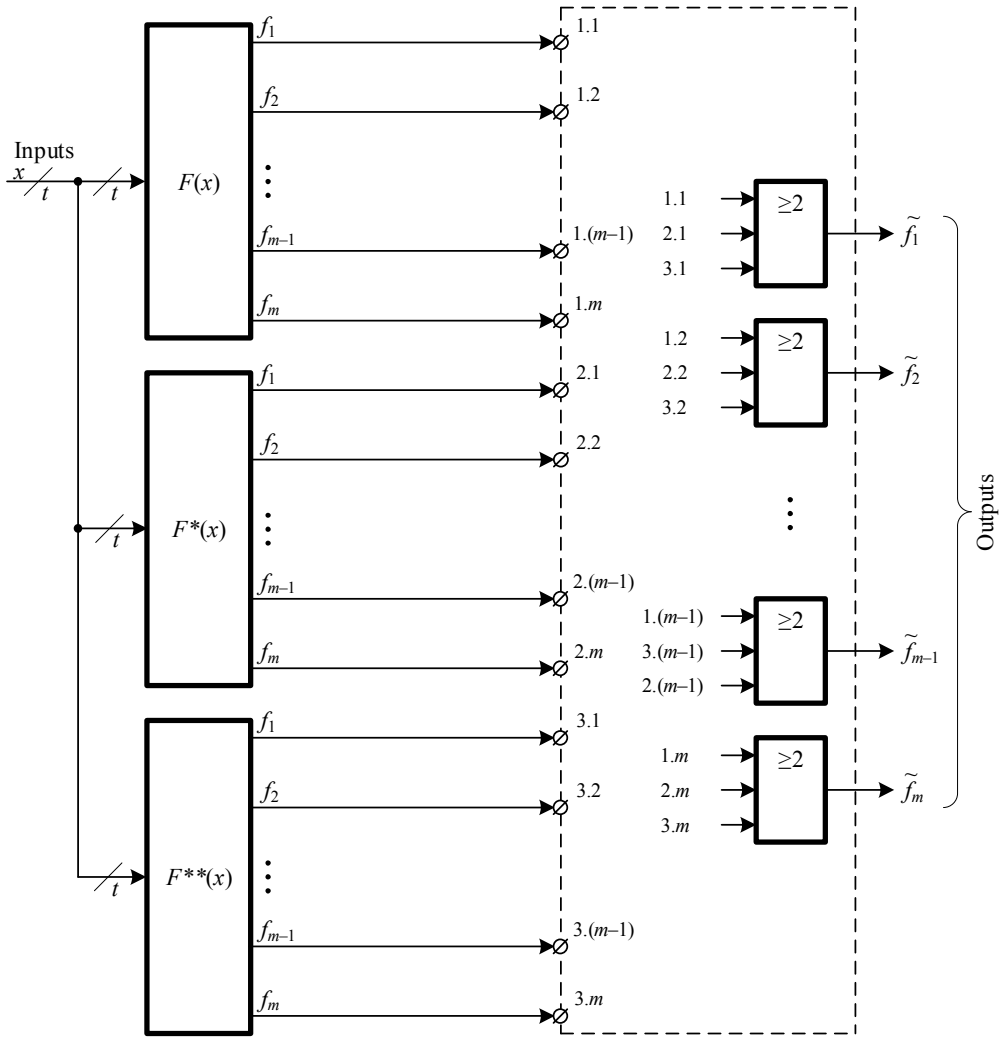


Fig. 1. *M*-structure of the signal correction.

The research shows that to create a structure with the signal correction, the principle of double modular redundancy can be applied with the control of a source device copy according to some diagnostic feature. This makes it possible to synthesize simpler devices that are insensitive to single faults and errors on circuit lines.

III. The Duplication Based Signal Correction Circuit with the Integrated Control Circuit

A. The generalized structure of the signal correction with the integrated control circuit

The duplication based signal correction structure is shown in the Fig. 2. This structure uses the original block $F(x)$, as well as its copy $F^*(x)$. The values at the same-name outputs of both blocks are compared at the inputs of the cascade of two-input elements of addition by modulo two. If there is a discrepancy in the values at the inputs of the element of addition by modulo two, a logical unit signal is generated at its output. It serves as an e_i error signal for each $i \in \{1,2,...,m\}$ of the device's $F(x)$ outputs. To exclude the correction for errors at the outputs of the $F^*(x)$ block, it is provided with a control circuit based on some diagnostic feature. The output z of the control circuit is connected to the input cascade of the correction circuit. It is formed by two-input logical multiplication elements that are set for each output of signal comparison elements from blocks $F(x)$ и $F^*(x)$. The first inputs of the logical multiplication elements receive signals from the comparison elements, and the second inputs receive a z signal from the control circuit of the $F^*(x)$ block. The value of the latter is inverted to eliminate false signal correction. This is necessary because the control circuit fixes the presence of errors exactly in the source device copy. The correction of functions calculated by the $F(x)$ block occurs at the inputs of the elements of addition by modulo two of the output cascade of the signal correction circuit.

The control circuit of the $F^*(x)$ block can be constructed, for example, on the basis of belonging of code vectors to separable (m,k)-codes, where m and k are the number of data and check bits. In this case, the outputs of the $F^*(x)$ block are connected to the inputs of the encoder of (m,k)-code $H(f)$, which forms the check vector $<h'_k\ h'_{k-1}\ \dots\ h'_2\ h'_1>$. A similar device $H(x)$, however, operating on the values of input influences of the $F(x)$ and $F^*(x)$ devices, forms an alternative check vector $<h_k\ h_{k-1}\ \dots\ h_2\ h_1>$. The vectors $<h'_k\ h'_{k-1}\ \dots\ h'_2\ h'_1>$ and $<h_k\ h_{k-1}\ \dots\ h_2\ h_1>$ are bitwise compared using the cascade of two-inputs adders by modulo two. If there is a discrepancy between the values at the inputs, the single values of the signals at the outputs of these elements are set. The outputs of the elements of addition by modulo two are connected to the inputs of the logical multiplication elements. The presence of a single value at the output of the multiplication element indicates the presence of an error at the outputs of the $F^*(x)$ block or in its control circuit.

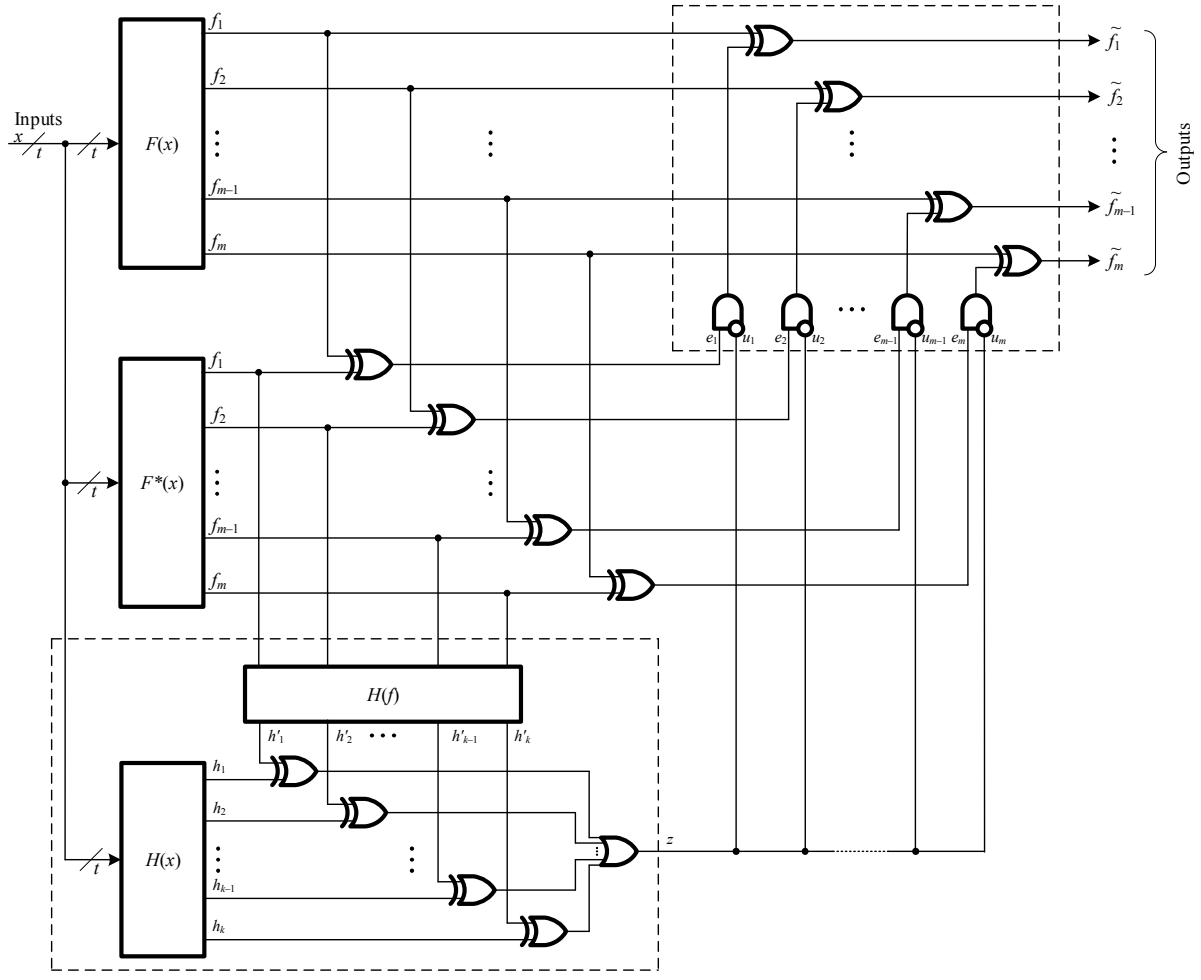


Fig. 2. The duplication based signal correction structure with the integrated control circuit.

This arrangement of the correction circuit in many cases makes it possible to synthesize simpler discrete devices that are insensitive to single faults than in the *M*-structure. It should be noted that the variants with optimization of the $F^*(x)$ block, as well as both $F(x)$ и $F^*(x)$ blocks are possible in the construction of structure presented in the Fig. 2, just like in the *M*-structure. This makes it possible to obtain a duplication structure with signal correction, which has minimal redundancy.

Let's consider three typical variants of the signal correction structure synthesis based on duplication with the control circuit. These variants are obtained through the use of the control circuits by the repetition codes [22], by the parity codes [23] and by a special code with summation of weighted transitions [24].

B. The structure of the signal correction with the integrated control circuit by the repetition codes

One way to implement a block diagram with double modular redundancy is to control a source device copy ($F^*(x)$ block) based on the repetition codes (Fig. 3). In this case, another source device copy is required (the $F^{**}(x)$ block). The presented signal correction circuit is called the *D*-structure. The organization of the control circuit by the repetition codes makes it possible to detect any faults in the controlled object that appear at its outputs. However, in fact, using repetition codes leads to a return to triplication (let's compare the Fig. 1 and the Fig. 3). The correction is carried out without using of majority elements in the *D*-structure shown in the Fig. 3, in contrast to the *M*-structure.

C. The structure of signal correction with the integrated control circuit by the parity code

In practice, it may be effective to use another typical structure of the signal correction circuit. It is based on the application of the control by the parity code (Fig. 4). The outputs of the $F^*(x)$ block are checked by convolution by modulo two, which implements the function $p' = f_1 \oplus f_2 \oplus \ldots \oplus f_m$. The value of this function is fed to the first input of the comparison element (addition by modulo two). The $P(x)$ block calculates the value of the parity function p on the values of the input influences of the $F(x)$ and $F^*(x)$ blocks. If the latter match, a control signal z=0 is generated. If the input values differ, an error signal is generated, that deactivates the correction circuit.

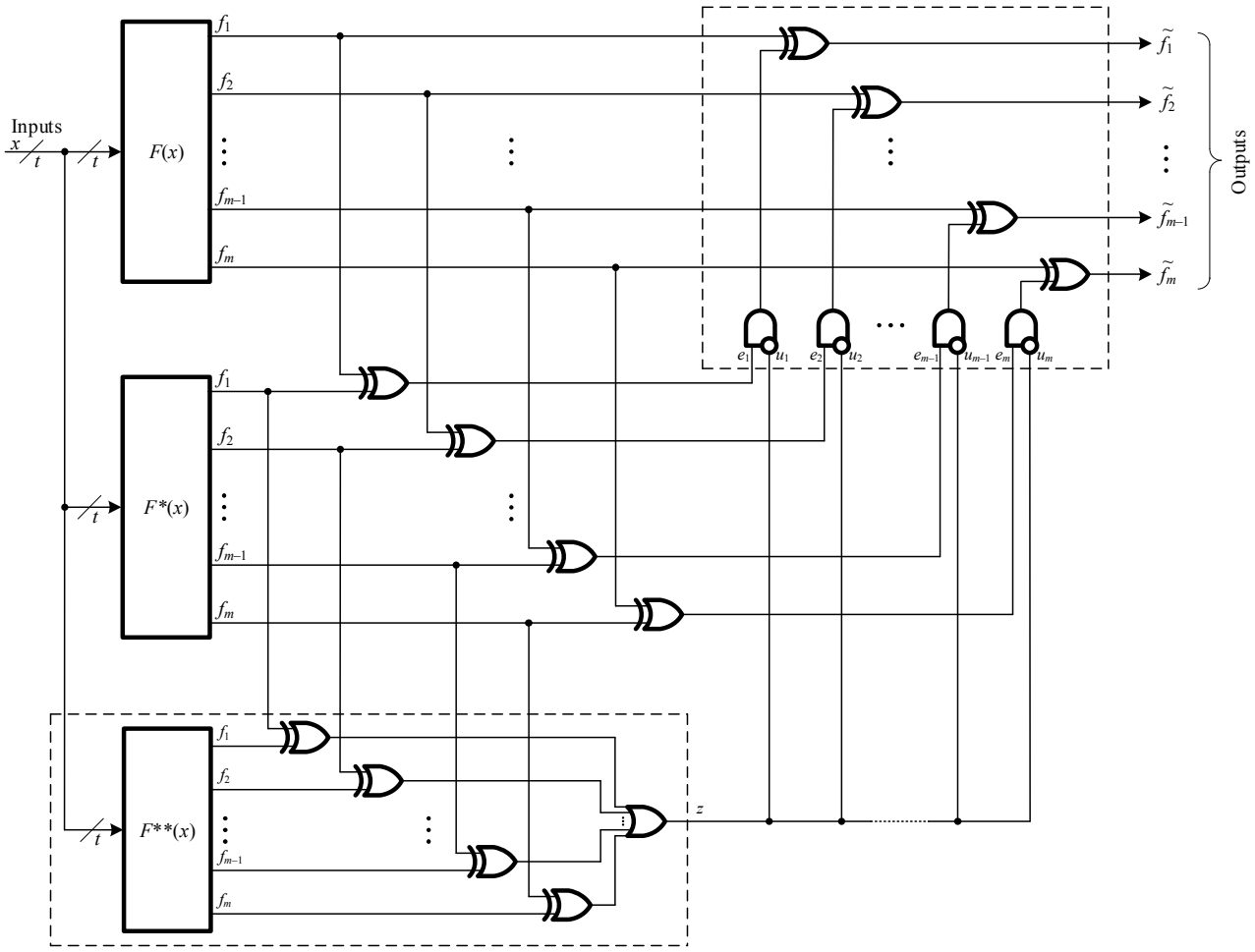


Fig. 3. *D*-structure of the signal correction.

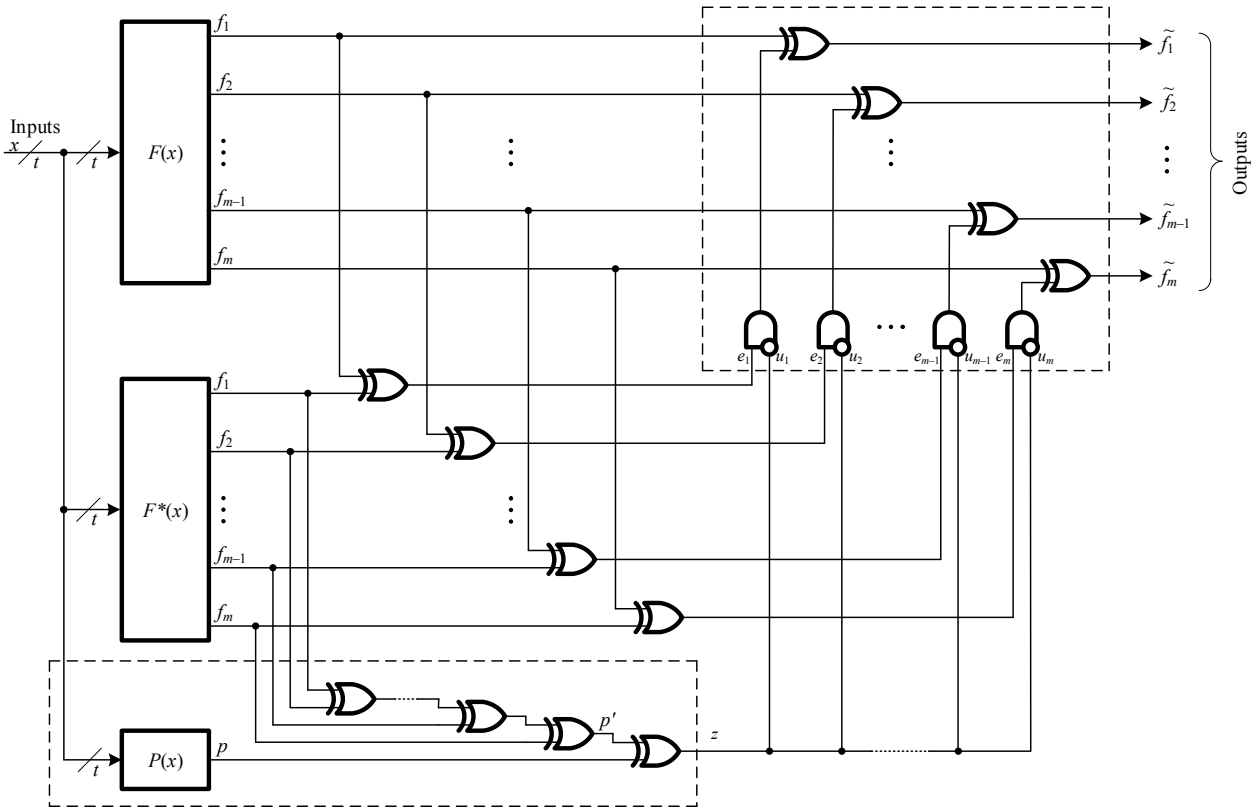


Fig. 4. *P*-structure of the signal correction.

The signal correction structure based on the double modular redundancy with the control of the calculation by parity is called the *P*-structure. Its advantage is that the control circuit is much simpler than if we use the repetition codes [2]. Due to this, in a large number of cases, it is possible to significantly reduce the complexity of the technical implementation of the *P*-structure in comparison with the *M*-structure. The disadvantage of the *P*-structure is the inability to detect the manifestations of any malfunctions at the outputs of the $F^*(x)$ block. The parity control circuit does not detect any error with an even multiplicity. This leads to the possibility of the false correction of signals at the device outputs when errors occur at the outputs of the $F^*(x)$ block and when the $F^*(x)$ block is correct. Nevertheless, there are known the methods for the control circuits synthesizing by the groups of independent outputs with their control by parity [25], as well as the methods for the converting of the device structures into the devices whose outputs form a single group of independent outputs [26]. The control by the groups of independent outputs and the control of a single group of independent outputs for most devices provides less redundant circuits than if we use the control by the duplication method.

D. The structure of the signal correction with the integrated control circuit by a special code with summation of weighted transitions

Any known code from a variety of the sum codes and their modifications can be used as a code for the $F^*(x)$ device checking [8]. The special code with summation of weighted transitions stands out among all sum codes according to its properties [24]. It has almost double redundancy $k=m-1$ and detects any errors with the exception of errors with the multiplicities $d=m$. At the same time, however, this code has simple control functions described by the convolutions by modulo two.

This code is designed according to the following rules:

1. The weight coefficients $w_{i,i+1}$ from a series of increasing powers of 2 are assigned to transitions between the bits of the data vector, starting from the lowest bit: $[w_{i,i+1}] = [w_{m-1,m}, w_{m-2,m-1}, \ldots, w_{2,3}, w_{1,2}] = [2^{m-1}, 2^{m-2}, \ldots, 2^1, 2^0]$.
2. The total weight of active transitions is calculated:

$$W = \sum_{i=1}^{m-1} w_{i,i+1} t_{i,i+1}, \tag{1}$$

where $t_{i,i+1} = f_i \oplus f_{i+1}$ is the function for activating the transition between the f_i and f_{i+1} bits.

3. The resulting number is presented in the binary form and is recorded in the bits of the check vector.

Let's denote the code with summation of weighted transitions as the $T(m,k)$-code, where m and k are the lengths of the data and check bits, respectively. As noted above, the $T(m,k)$-code has the $k=m-1$ check bit. The values of the check bits can be determined by the formulas:

$$\begin{aligned} h_1 &= f_1 \oplus f_2; \\ h_2 &= f_2 \oplus f_3; \\ &\ldots \\ h_{m-1} &= f_{m-1} \oplus f_m. \end{aligned} \tag{2}$$

To obtain the values of the check bits of the code with the summation of weighted transitions, only the addition operations by modulo two are used, therefore, the encoder structure of this code will be standard and contain the $m-1$ element of addition by modulo two. The presence of a

standard encoder structure makes it possible to synthesize a typical error correction structure (Fig. 5). Let's call this structure a *T*-structure.

The *T*(*m*,*k*)-code will detect any distortions in the checked code vector, with the exception of errors with a maximum multiplicity of *d*=*m*. This is because the value of the total weight of the data vector calculated using the formula (1) will not change only if it is calculated for two vectors with completely opposite bit values. This feature of the *T*(*m*,*k*)-code makes it possible to use it very effectively for organizing control of combinational logic devices. At the same time, only one restriction is imposed on the structures of the checked devices: there are no paths from any internal logical elements leading immediately to all their outputs (we can say that this is a structural restriction). However, the condition of the impossibility of simultaneous distortion of all *m* outputs of the device can be checked even if there are such elements [27]:

$$\frac{\partial f_1}{\partial y_q} \cdot \frac{\partial f_2}{\partial y_q} \cdot \ldots \cdot \frac{\partial f_m}{\partial y_q} = 0, \qquad (3)$$

where y_q is a function that is implemented at the output of a logical element G_q that is connected by paths to all outputs of the device.

If the expression on the left side of formula (3) is equal to zero, then there is no input set on which errors are transmitted to all outputs of the device.

E. The classification of the signal correction structures

The error correction structures presented above, based on double modular redundancy with the control of calculation, makes it possible to propose new methods for the synthesis of fault-tolerant discrete systems (Fig. 6). These structures should be taken into account when choosing a method for implementing a device or a system that is insensitive to single component failures.

IV. Experimental results

In the research of the characteristics of signal correction circuits, the authors conducted experiments to evaluate the effectiveness of each of the proposed structures. The experiments included two stages. The first stage of the experiments was to assess the complexity of the technical implementation of *D*-, *P*- and *T*-structures for a set of control combinational circuits from the MCNC Benchmarks database. The second stage of the experiments was to evaluate the characteristics of error detection at the outputs of combinational circuits from the LGSynth'89 database [28].

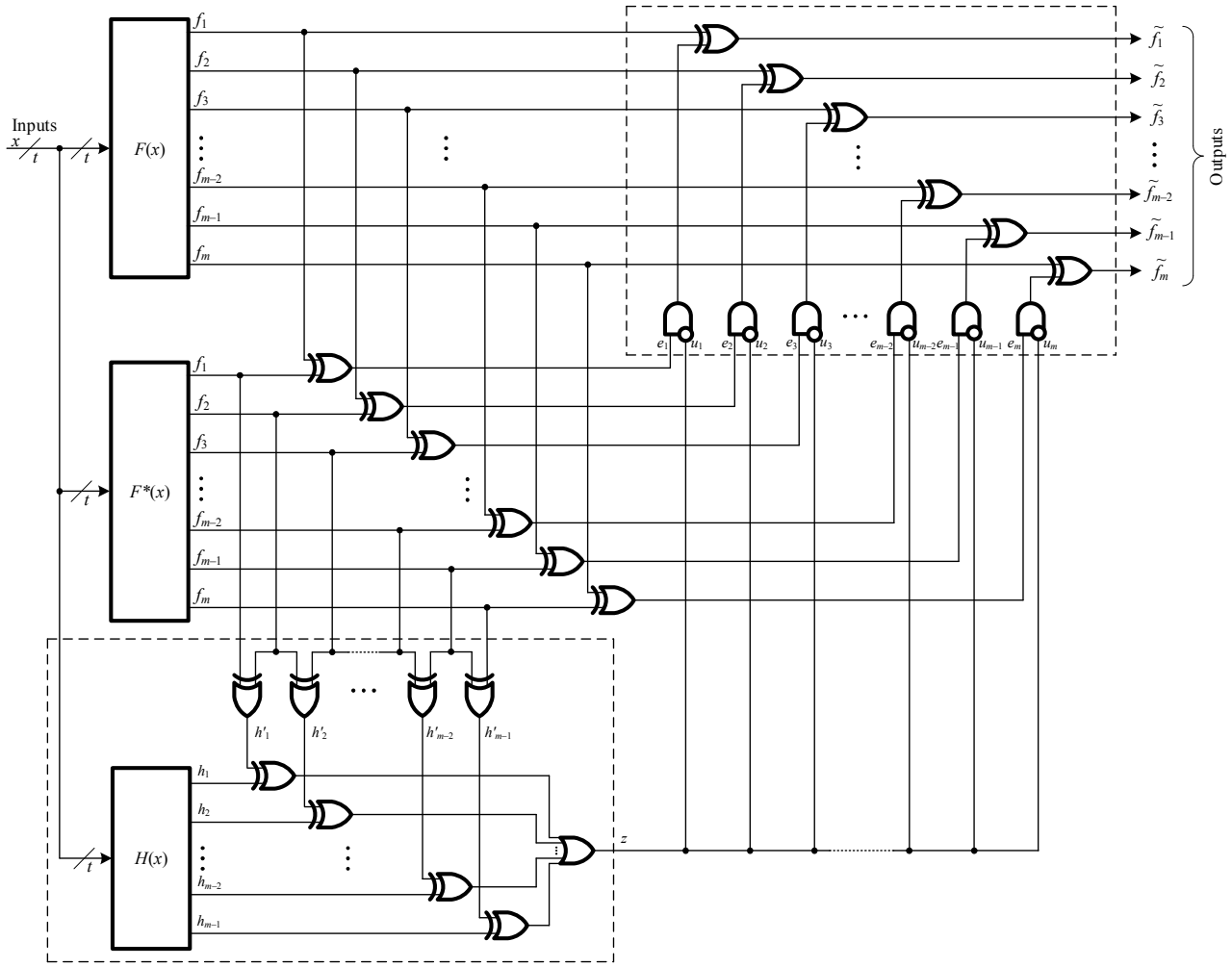


Fig. 5. *T*-structure of the signal correction.

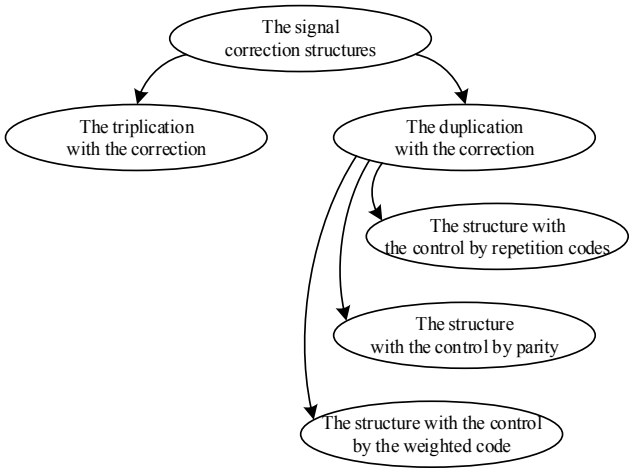


Fig. 6. The classification of the signal correction structures.

At the first stage, the signal correction structures considered above were constructed for control combinational circuits. We determine the area occupied by each structure (in conventional units), taking into account the use of the library of functional elements *stdcell*2_2.*genlib* [29]. This made it possible to obtain data on the areas of four structures: *M*-, *D*-, *P*- и *T*-structures of signal correction. Each of the proposed structures based on double modular redundancy was compared with the structure of triple modular redundancy. We determined the ratio indicator of the share of the proposed correction circuit area to the area of the *M*-structure:

$$\delta = \frac{L_D}{L_M} \cdot 100\%, \quad \tau = \frac{L_T}{L_M} \cdot 100\%, \quad \pi = \frac{L_P}{L_M} \cdot 100\%. \quad (4)$$

The obtained results are summarized in the Table 1, and also complemented by the graphs shown in the Fig. 7.

After analyzing the data in the Table 1, we obtain the following conclusions. For all considered circuits, the *D*-structure has a slightly larger implementation area than the *M*-structure. The average value of the indicator δ=102.65%. This shows that the *D*-structure as a whole is comparable in complexity to the *M*-structure. The *T*-structure using is more effective in terms of technical implementation complexity. We obtained simpler *T*-structures than *M*-structures for 18 of 25 combinational circuits. The average value of the indicator τ=94.614%. But we obtained the values of the indicator τ<90% for 9 combinational circuits. Taking into account the high features of error detection and correction, the result shows a significant advantage of the *T*-structure over the *M*-structure. The *P*-structure using provides the maximum reduction in the complexity of technical implementation in comparison with the *M*-structure among the considered structures of correction circuits. For the considered combinational circuits, the average value of the indicator π=80.612%.

To evaluate the characteristics of signal detection and correction, we conducted the experiments with modeling stuck-at faults at the outputs of internal logic elements of control combinational circuits. In the course of the experiment, we estimated the number of undetectable errors that occur during all single faults are sequentially introduced into the circuit structure when all input combinations are applied to its inputs. The achieved results are listed in the Table 2. The experiments confirm the theoretical research of the authors. A certain number of errors in the *P*-structure is not detected for the majority of circuits (and therefore correction is not possible). This is 10.553% of all possible errors on average. All errors were detected for the three circuits. Less than 10% of all errors are not detected for another seven circuits. Any errors are detected and corrected in the *T*-structure constructed for 19 of 21 circuits. All errors cannot be identified for two circuits whose structures allow distortions with multiplicities *d*=*m*. However, the percentage of such errors is extremely small and amounted to less than 1%.

The results obtained in the course of experimental researches of new signal correction structures indicate the high efficiency of the two proposed structures: *P*- and *T*-structures. Their use in practice can provide with simpler circuits of the fault-tolerant devices and systems than if we use the traditional correction structure according to the majority principle.

V. Conclusion

In the synthesis of fault-tolerant discrete devices and systems, structures based on double modular redundancy can be used instead of the traditional correction structure with triple modular redundancy (*M*-structure). In this case, a source device copy should be provided with a control circuit for some diagnostic feature. In the research of the authors, it is proposed to perform the control by the separable repetition codes, by the parity codes and by the special sum codes. This makes it possible to synthesize typical structures of signal correction circuits.

The analysis of the *D*-, *P*- and *T*-structures of signal correction circuits proposed by the authors showed the following results. Any errors can be corrected in the *D*-structure, as well as in the *M*-structure. However, the *D*-structure is slightly less complex than the *M*-structure. In the experiment, for all test examples, the value of the technical implementation complexity indicator was obtained for *D*-structures greater than for *M*-structures. The excess, however, is insignificant – for many circuits it is no more than 2-3%. The average value of the share of the area occupied by the *D*-structure from the area occupied by the *M*-structure made up 102.65%. For most test examples, smaller area values were obtained for the *T*-structure than for the *M*-structure. The average value of the share of the area occupied by the *T*-structure from the area occupied by the *M*-structure made up 94.614%. At the same time, any error that does not cause distortion of all outputs of the source device copy is detected and corrected in the *T*-structure. The indicators of the complexity of the *P*-structure implementation show that this correction structure is the simplest of the proposed ones. The average value of the share of the area occupied by the *P*-structure from the area occupied by the *M*-structure made up 80.612%. The disadvantage of the presented correction structure is a much lower correction capacity. Any failure with an even multiplicity of the source device copy is not detected in the *P*-structure. This does not make it possible to correct any errors in the source device. However, special circuitry methods can be used for implementing control circuits that increase the detecting and correcting capabilities of the *P*-structure.

TABLE I. EXPERIMENTAL RESULTS ON ASSESSING THE COMPLEXITY OF TECHNICAL IMPLEMENTATION OF SIGNAL CORRECTION STRUCTURES

No.	$F(x)$	$L_{F(x)}$	L_M	L_D	L_T	L_P	δ, %	τ, %	π, %
1	b2	40952	125032	125592	108760	87624	100.448	86.986	70.081
2	br1	3608	11848	12120	11192	9048	102.296	94.463	76.367
3	br2	2952	9880	10152	9152	7608	102.753	92.632	77.004
4	dc1	976	3824	4064	3800	3120	106.276	99.372	81.59
5	dekoder	736	3104	3344	3320	2728	107.732	106.959	87.887
6	dist	6968	21544	21720	19360	17416	100.817	89.863	80.839
7	gary	10688	33472	33840	34904	25440	101.099	104.278	76.004
8	in0	10704	33520	33888	34936	25472	101.098	104.224	75.99
9	in1	40952	125032	125592	108760	87624	100.448	86.986	70.081
10	inc	2376	8280	8584	8608	6560	103.671	103.961	79.227
11	intb	22248	67640	67880	72072	96160	100.355	106.552	142.164
12	m1	3064	10728	11128	9936	8160	103.729	92.617	76.063
13	m2	10096	32336	32864	26968	23240	101.633	83.399	71.87
14	m3	13464	42440	42968	34888	30368	101.244	82.205	71.555
15	m4	18704	58160	58688	48520	41472	100.908	83.425	71.307
16	max512	9632	29664	29872	25816	22688	100.701	87.028	76.483
17	max1024	17816	54216	54424	47184	41392	100.384	87.03	76.346
18	mlp4	7224	22696	22968	22432	17936	101.198	98.837	79.027
19	newcpla2	1896	6968	7304	6864	5680	104.822	98.507	81.515
20	newcwp	440	1960	2136	2032	1800	108.98	103.673	91.837
21	newtpla2	840	3032	3176	2856	2448	104.749	94.195	80.739
22	p82	2368	8896	9360	9160	7160	105.216	102.968	80.486
23	root	3496	11128	11304	9624	8832	101.582	86.485	79.367
24	sqn	2008	6408	6520	6272	5672	101.748	97.878	88.514
25	tms	6784	22400	22928	20344	16344	102.357	90.821	72.964
Average		9639.68	30168.32	30496.64	27510.4	24079.68	102.65	94.614	80.612

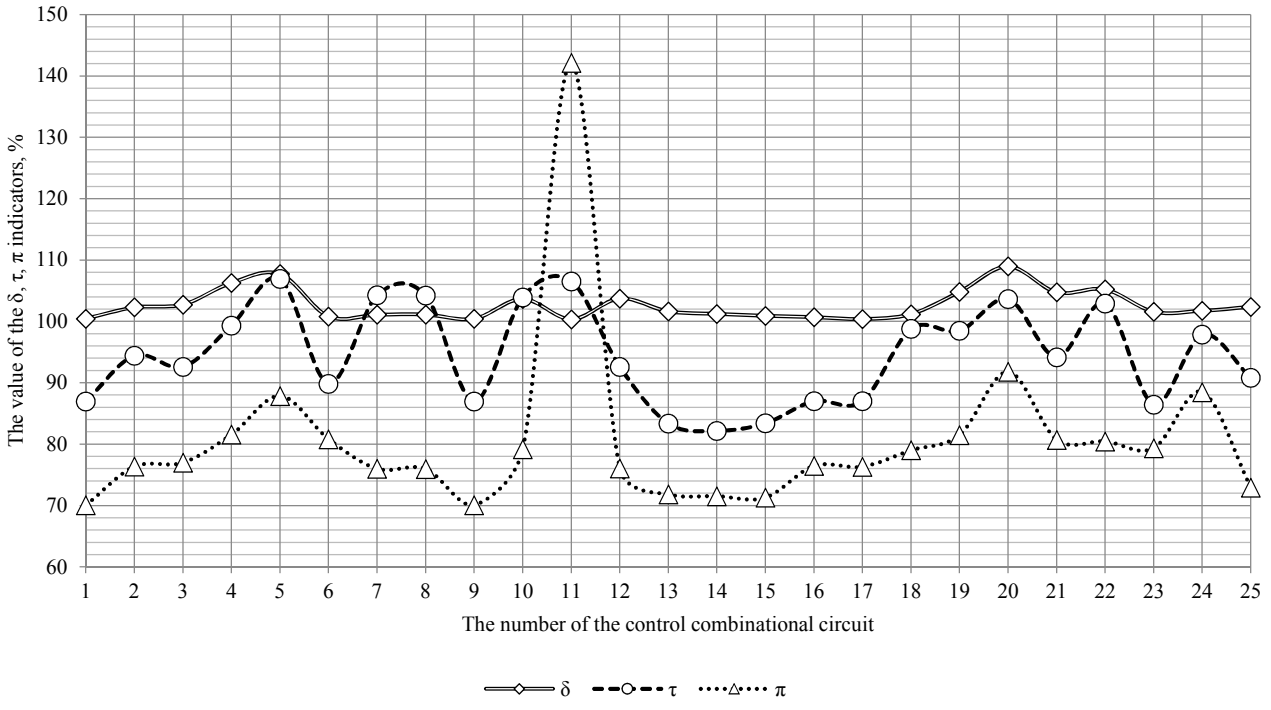


Fig. 7. The graphical representation of indicators of the complexity of the *D*-, *T*- and *P*-structures implementation.

TABLE II. EXPERIMENTAL RESULTS ON ASSESSING THE CHARACTERISTICS OF ERROR DETECTION IN THE SIGNAL CORRECTION STRUCTURES

No.	$F(x)$	The number of inputs/ outputs	The number of undetectable errors on the outputs of the $F^*(x)$ block in typical structures			The total number of errors on the outputs	The share of undetectable errors from their total number, %		
			M-	*P*-	*T*-		*M*-	*P*-	*T*-
1	z4ml	7 / 4	0	128	0	4168	0	3.071	0
2	b1	3 / 4	0	2	0	46	0	4.348	0
3	cmb	16 / 4	0	39462	0	288218	0	13.692	0
4	cm162a	14 / 5	0	44763	224	317331	0	14.106	0.071
5	cm163a	16 / 5	0	153920	64	1221312	0	12.603	0.005
6	alu2	10 / 6	0	12663	0	62838	0	20.152	0
7	x2	10 / 7	0	2524	0	19708	0	12.807	0
8	alu4	14 / 8	0	372633	0	1980377	0	18.816	0
9	cm138a	6 / 8	0	0	0	680	0	0	0
10	f51m	8 / 8	0	887	0	13264	0	6.687	0
11	pcle	19 / 9	0	1018583	0	17472087	0	5.83	0
12	cm42a	4 / 10	0	8	0	278	0	2.878	0
13	cu	14 / 11	0	61888	0	137984	0	44.852	0
14	pm1	16 / 13	0	43776	0	757760	0	5.777	0
15	sct	19 / 15	0	557008	0	16586128	0	3.358	0
16	decod	5 / 16	0	0	0	224	0	0	0
17	tcon	17 / 16	0	0	0	4849664	0	0	0
18	pcler8	27 / 17	0	917294976	0	4331229952	0	21.179	0
19	ldd	9 / 19	0	4813	0	30182	0	15.947	0
20	cc	21 / 20	0	3873752	0	35167192	0	11.015	0
21	ttt2	24 / 21	0	33948368	0	755063504	0	4.496	0

It should be noted that all the structures proposed in this paper are typical. Each of the structure components is present in any database of functional elements of computer-aided design tools. The control logic blocks of control circuits are obtained by optimizing two-cascade circuits: the first cascade is the source device itself, the second cascade is the encoder of the code used in the correction structure.

In all considered structures, including the *M*-structure, errors of the correction circuits are not corrected directly. In structures based on double modular redundancy, these are errors of elements of logical multiplication and addition by modulo two. In the structure based on triple modular redundancy, these are errors of the output cascade of the majority elements. This problem is solved by using a highly reliable elemental base in the implementation of the correction circuits themselves.

The usage of structures based on double modular redundancy makes it possible in some cases to synthesize simpler fault-tolerant discrete devices and systems with high correcting capacity than structures based on triple modular redundancy.

REFERENCES

[1] N.S. Scherbakov "The Reliability of Digital Devices" (in Russ.), Moscow, Mechanical Engineering, 1989, 224 p.

[2] E.S. Sogomonyan, and E.V. Slabakov "Self-Checking Devices and Fault-Tolerant Systems" (in Russ.), Moscow: Radio & Communication, 1989, 208 p.

[3] D.V. Gavzov, V.V. Sapozhnikov, and Vl.V. Sapozhnikov "Methods for Providing Safety in Discrete Systems", Automation and Remote Control, 1994, vol. 55, issue 8, pp. 1085-1122.

[4] E.S. Sogomonyan "Self-Correction Fault-Tolerant Systems", Preprint, October 2018, 30 p.

[5] W.E. Ryan, and S. Lin "Channel Codes: Classical and Modern", Cambridge University Press, 2009, 708 p.

[6] E. Fujiwara "Code Design for Dependable Systems: Theory and Practical Applications", John Wiley & Sons, 2006, 720 p.

[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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] P.K. Lala "Self-Checking and Fault-Tolerant Digital Design", San Francisco: Morgan Kaufmann Publishers, 2001, 216 p.

[13] V.V. Sklyar, and V.S. Kharchenko "Fault-Tolerant Computer-Aided Control Systems with Multiversion-Threshold Adaptation: Adaptation

Methods, Reliability Estimation, and Choice of an Architecture", Automation & Remote Control, 2002, Vol. 63, Issue 6, pp. 991-1003, doi: 10/1023/A:1016130108770.

[14] M. Hamamatsu, T. Tsuchiya, and T. Kikuno "Finding the Optimal Configuration of a Cascading TMR System", 14th IEEE Pacific Rim International Symposium on Dependable Computing,15-17 December 2008, Taipei, Taiwan, pp. 329-350, doi: 10.1109/PRDC.2008.12.

[15] A. Chakraborty "Fault Tolerant Fail Safe System for Railway Signalling", Proceedings of the World Congress on Engineering and Computer Science (WCECS 2009), USA San Francisco, Vol. II, October 20-22, 2009.

[16] K. Matsumoto, M. Uehara, and H. Mori "Evaluating the Fault Tolerance of Stateful TMR", 13th International Conference on Network-Based Information Systems, 14-16 September 2010, Takayama, Japan, pp. 332-336, doi: 10.1109/NBiS.2010.86.

[17] J. Borecký, M. Kohlík, P. Vít, and H. Kubátová "Enhanced Duplication Method with TMR-Like Masking Abilities", Euromicro Conference on Digital System Design (DSD), 31 August – 2 September 2016, Limassol, Cyprus, pp. 690-693, doi: 10.1109/DSD.2016.91.

[18] 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.

[19] G. Theeg, and S. Vlasenko "Railway Signalling & Interlocking: 3ed Edition", Germany, Leverkusen PMC Media House GmbH, 2020, 552 p.

[20] V.V. Sapozhnikov, Vl.V. Sapozhnikov, and D.V. Efanov "Reliability and Technical Diagnostics Theory Fundamentals" (in Russ.), St. Petersburg: «Lan′» Pub. House, 2019, 588 p.

[21] M. Đug, S. Weidling, E.S. Sogomonyan, D. Jokic, and M. Krstic "Full Error Detection and Correction Method Applied on Pipelined Structure Using Two Approaches", Journal of Circuits, Systems and Computers, 17 January 2020, pp. 1-16, doi: 10.1142/S0218126620502187.

[22] N.T. Berezyuk, A.G. Andrushchenko, S.S. Moshchickij, V.I. Glushkov, M.M. Bekesha, and V.A. Gavrilov "Information Encoding (Binary Codes)" (in Russ.), edited N.T. Berezyuk, Kharkov: «Vishcha shkola», 1978, 252 p.

[23] 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.

[24] V.V. Sapozhnikov, Vl.V. Sapozhnikov, D.V. Efanov, and V.V. Dmitriev "New Structures of the Concurrent Error Detection Systems for Logic Circuits", Automation & Remote Control, 2017, vol. 78, issue 2, pp. 300-313, doi: 10.1134/S0005117917020096.

[25] M. Goessel, A.A. Morozov, V.V. Sapozhnikov, and Vl.V. Sapozhnikov "Investigation of Combination Self-Testing Devices Having Independent and Monotone Independent Outputs", Automation & Remote Control, 1997, Vol. 58, Issue 2, pp. 299-309.

[26] A. Morosow, V.V. Sapozhnikov, Vl.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.

[27] D.V. Efanov, V.V. Sapozhnikov, and Vl.V. Sapozhnikov Sapozhnikov Synthesis of Self-Checking Combination Devices Based on Allocating Special Groups of Outputs", Automation and Remote Control, 2018, Vol. 79, Issue 9, pp. 1609-1620, doi: 10.1134/S0005117918090060.

[28] Collection of Digital Design Benchmarks [http://ddd.fit.cvut.cz/prj/Benchmarks/].

[29] 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.

Kuramoto Model for Oscillators with Fractional Frequencies Ratios in Circuit Analysis Application

Mark M. Gourary,
IPPM, Russian Academy of Sciences
Moscow, Russia
rusakov@ippm.ru

Sergey G. Rusakov, *Member IEEE*
IPPM, Russian Academy of Sciences
Moscow, Russia
rusakov@ippm.ru

***Abstract*— The problems of joint application of Kuramoto model and circuit simulation algorithms are considered. The method to evaluate the model parameters for practical oscillator circuits is proposed. The parameters of oscillators are represented by harmonics of their periodic steady-state solutions and perturbation projection vectors. The parameters of couplings are defined by their frequency dependent transfer functions. The developed approach for oscillators with close frequencies is extended to the general case when the ratios of the oscillators' natural frequencies are close to rational fractions.**

Keywords—coupled oscillators, circuit simulator, phase macromodel, fractional frequencies ratios

I. INTRODUCTION

The analysis of oscillatory interactions in electronic circuits [1, 2] is the important problem in the design of oscillators in radiofrequency integrated circuits (RFICs). The aim of the analysis is the evaluation of perturbations caused by the oscillator's interaction with other IC blocks including mutual coupling of oscillators. The application of circuit simulation tools is effective for a small number of oscillators but it requires too high computational efforts for large networks. Implementation of simplified techniques to solve this problem is the way of reducing the computational efforts.

Kuramoto model (KM) [3] is specified by system of Ordinary Differential Equation (ODE). This is the most popular macromodel in different research applications. The following areas of applying KM can be mentioned: biology, medicine, chemistry, social sciences, neural networks [4, 5], etc. Each system variable of KM represents the phase of the corresponding oscillator.

However, the use of the KM for the analysis of coupled electronic oscillators meet the following significant difficulties.

1. KM is a phenomenological model that is usually applied to a qualitative analysis of oscillatory systems. The KM parameters are not determined by the characteristics of the oscillatory circuit and cannot be obtained by its simulation.

2. Oscillators' interactions in most cases are characterized by constant coupling strength. To account for the dynamic properties some particular approaches were presented [6-8]. However, there are no approaches taking into account dynamic characteristics when the coupling strengths are set by transfer functions of electrical networks.

3. KM describes an oscillatory system with close values of oscillators' natural frequencies. However, synchronization effects can also arise if ratios of natural frequencies are close to rational fractions [9]. Such effects can appear in IC due to the large number of different types of oscillators on chip. Thus, the analysis of parasitic interactions in IC requires applying the model that covers the cases of rational ratios of frequencies.

In this paper we propose approach eliminating pointed above shortcomings of KM. The paper is organized as follows. Section 2 describes equations of known phase macromodels including KM. In Section 3 we derive parametrized KM by the application of smoothed macromodel equations to the analysis of oscillators connected by linear dynamic systems. Section 4 contains the development of parametrized KM for oscillators with natural frequencies close to rational fractions.

II. MATHEMATICAL REPRESENTATION OF KNOWN OSCILLATOR MODELS

A. Kuramoto Model

The following ODE system represents KM general form [3] for N coupled oscillators

$$\frac{d\theta_m}{dt} = \omega_m + \sum_{n=-N}^{N} u_{mn}(\theta_m - \theta_n), \qquad m = 1 \ldots N, \quad (1)$$

Here ω_m, θ_m define fundamental of m-th oscillator and its instantaneous phase correspondingly, $u_{mn}(\theta_{mn})$ are 2π-periodic coupling functions, $\theta_{mn}=\theta_m - \theta_n$.

The natural extension of KM also considers the external periodic force [3] as an additional oscillator $\omega_{N+l}, \theta_{m,N+l}$ unidirectionally coupled with internal ones. In such case (1) is extended by external coupling functions of the form $u_{m,N+l}(\theta_{N+l} - \theta_m)$ included in the Right-Hand Side (RHS) of m-th equation. The number of equations and the set of its variables $\theta_1, \ldots, \theta_N$ are saved because the excitation phases θ_{N+l} are known.

In the simplest case when one oscillator is excited by a single stimulus the KM equation has the form

$$\frac{d\theta}{dt} = \omega_0 + u_{ex}(\theta - \theta_{ex}). \quad (2)$$

The definitions of coupling functions $u_{mn}(\theta mn)$ in KM do not include the characteristics of oscillators and interconnects (1) that is essential shortcoming of the model.

The study was funded by RFBR project no. 19-29-03012.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

B. PPV Micromodel

An approach to analyze oscillators' ensemble considering its real characteristics is presented by the nonlinear phase macromodel [10] that is often called as PPV macromodel. It exploits the following characteristics of an excited oscillator:

- $x(t)$ is the periodic steady state (PSS) solution of the ODE system for the free running oscillator,
- $v(t)$ is the periodic Perturbation Projection Vector (PPV) also computed from the ODE system [11],
- $b(t)$ is the small vector of excitations

The solution of the excited oscillator ODE is assumed to be presented as $x_p(t) = x(t + \alpha(t))$ with the time-varying delay $\alpha(t)$ is obtained by the nonlinear ODE [10, 11]:

$$\frac{d\alpha(t)}{dt} = v\big(t + \alpha(t)\big) \cdot b(t), \alpha(0) = 0. \quad (3)$$

ODE (3) can be formed for each oscillator of the ensemble taking into account that PPV is the oscillator property, and the excitation $b(t)$ is defined by the waveform of other oscillator and interconnect characteristics. This approach to analyze the network of N coupled oscillators was proposed in [12]:

$$\frac{d\alpha_m(t)}{dt} = v_m^T\big(t + \alpha_m(t)\big) \cdot \sum_{n=-N}^{N} \gamma_{mn}(t), \ \ \alpha(0) = 0. \quad (4)$$

Here $\gamma_{mn}(t)$ is the excitation of m-th oscillator produced by the waveform $x_n\big(t + \alpha_n(t)\big)$ of n-th oscillator through the mn interconnect defined by linear time invariant (LTI) system [12]

$$C_{mn}\frac{dz}{dt} + G_{mn}z = F_{mn}x_n\big(t + \alpha_n(t)\big), \ \gamma_{mn} = D_{mn}z. \quad (5)$$

The variable z in [5] is internal state vector of the LTI system with matrices G_{mn}, C_{mn}, F_{mn}, D_{mn}. These matrices represent basic parameters that characterize the dynamic properties of IC blocks when their signals propagate through IC interconnects. The behavior of coupled oscillators is determined by joint solving of (4, 5).

As shown in [13] the solution (3) includes high-frequency oscillations that slow down the simulation and make it difficult to determine the synchronization conditions. These shortcomings also persist in the extended system (4), (5).

C. Smoothed PPV Macromodel

To eliminate shortcomings of the time-domain PPV micromodel (3) its smoothed version was derived in [13] as the phase equation the frequency domain. The equation is applied to an excited oscillator with natural frequency ω_0. The small excitation has the form of Fourier series with slowly time-varying complex amplitudes defined by the vector $\tilde{B}(t)$

$$b(t) = \sum_{k=-K}^{K} \tilde{B}_k(t)\, exp(jk\omega_0 t). \quad (6)$$

The oscillator waveform under the excitation is defined by the waveform of the unperturbed oscillator with harmonics X_k and the slowly time-varying phase shift $\varphi(t)$

$$x(t) = \sum_{k=-K}^{K} X_k\, exp\Big(jk\big(\omega_0 t + \varphi(t)\big)\Big). \quad (7)$$

Phase $\varphi(t)$ is obtained by solving ODE

$$\frac{1}{\omega_0}\frac{d\varphi}{dt} = W\Big(\tilde{B}(t), \varphi(t)\Big), \quad (8)$$

where $W(\tilde{B}, \varphi)$ is 2π-periodic function with respect to φ

$$W(\tilde{B}(t), \varphi) = \sum_{k=-K}^{K} \tilde{B}_k(t) V_k\, exp(-jk\varphi). \quad (9)$$

Here V_k is k-th harmonic of the oscillator's PPV $v(t)$.

For pure periodical excitation with frequency $\omega_0 + \Delta\omega$ and time-independent harmonics, B_k one can obtain the synchronization conditions by substituting $d\varphi/dt = \Delta\omega$ and $\tilde{B}_k(t) = B_k\, exp(jk\Delta\omega t)$ into (8), (9). The obtained algebraic equation with respect to the oscillator's locking phase φ is

$$\Delta\omega/\omega_0 = W(B, \varphi_0). \quad (10)$$

Maximal and minimal values of $W(B, \varphi)$ define the oscillator locking range $\Delta\omega$ for the given excitation waveform

$$\min_{0\le\varphi<2\pi} W(B, \varphi) \le \frac{\Delta\omega}{\omega_0} \le \max_{0\le\varphi<2\pi} W(B, \varphi). \quad (11)$$

This result was obtained earlier by direct analysis [14].

III. Derivation of Parametrized Kuramoto Model

In this section, we derive parametrized KM with dynamic couplings by the analysis of coupled oscillators using smoothed macromodel equations.

A. Representation of Kuramoto Model through the characteristics of free-running oscillators

Phase equation (8) jointly with (9) define the similar oscillator behavior as simple KM in (2) if the excitation is produced by the external oscillator waveform defined similarly with (7)

$$x^{ex}(t) = \sum_{k=-K}^{K} X_k^{ex}\, exp\, jk\big(\omega_{ex}t + \varphi_{ex}(t)\big). \quad (12)$$

We assume that produced excitation $b^{ex}(t)$ can be defined through its coupling factor with the external waveform (12)

$$b^{ex}(t) = K^{ex}x^{ex}(t), \quad (13)$$

After substituting (12) into (13) and comparing obtained terms with (6) one can conclude that

$$\tilde{B}_k(t) = K^{ex}X_k^{ex}\, exp\Big(jk\big(\Delta\omega_{ex}t + \varphi_{ex}(t)\big)\Big). \quad (14)$$

Substituting (14) into (9) leads to

$$W\big(\tilde{B}^{ex}, \varphi\big) = \sum_{k=-K}^{K} K^{ex}X_k^{ex}V_k\, exp\, jk(\varphi_{ex} + \Delta\omega_{ex}t - \varphi). \quad (15)$$

Relative phases φ, φ_{ex} of the oscillators are expressed through the corresponding instantaneous phases θ, θ_{ex} by

$$\varphi = \theta - \omega_0 t, \qquad \varphi_{ex} = \theta_{ex} - \omega_{ex}t \quad (16)$$

and the time derivative

$$d\varphi/dt = d\theta/dt - \omega_0, \quad (17)$$

After applying (16), (17) to replace the variables in (8), (15) and performing elementary transformations we obtain the equation (2) with the coupling function in the form

$$u_{ex}(\Delta\theta) = \sum_{k=-K}^{K} K^{ex}\omega_0 X_k^{ex} V_k \, exp(jk\Delta\theta). \quad (18)$$

Thus we obtained the representation of 2π-periodic coupling function $u_{ex}(\Delta\theta)$ by Fourier series (18) with harmonic magnitudes $U_k^{ex} = K^{ex}\omega_0 X_k^{ex} V_k$ that are defined by parameters of exciting oscillator (the fundamental ω_0 and harmonics of free-running oscillations X_k^{ex}) and parameters of the perturbed oscillator (PPV harmonics V_k).

This result can be easily expanded on the general form (1) of KM by defining coupling functions $u_{mn}(\Delta\theta)$ as Fourier series with the representation of the magnitude of *k*-th harmonic in the following view:

$$U_k^{mn} = K^{mn}\omega_m V_k^m X_k^n \quad (19)$$

Here K^{mn} is the coupling factor between the output waveform of *m*-th oscillator and the excitation waveform applied to *n*-th oscillator. Parameters ω_m, V_k^m, X_k^n can be evaluated by simulating the behavior of the corresponding oscillator in the free-running mode.

B. Kuramoto Model with Dynamic Frequency-Dependent Couplings

The assumption of a constant K^{ex} in (13) is not always true in practice especially in electronic circuits. More frequently the excitation is transferred through a linear interconnect with the transfer function (TF) in the frequency domain $H^{mn}(\omega)$. TF for LTI system is obtained by solving linear algebraic system and for (5) it is defined as

$$H^{mn}(\omega) = D^{mn}(G^{mn} + j\omega C^{mn})^{-1} F^{mn} \quad (20)$$

The waveform of *n*-th oscillator has the form (12). Taking into account (16) it is presented as

$$x^n(t) = \sum_{k=-K}^{K} X_k^n \, exp\big(jk\theta_n(t)\big). \quad (21)$$

The expression (21) represents oscillation with slow-varying instantaneous frequency $\omega_n^{inst} = d\theta_n/dt = \dot{\theta}_n$. The instantaneous frequency of *k*-th harmonic in (21) is $\dot{\theta}_n$. Thus, coupling factors between the harmonic of (21) and the corresponding excitation term can be evaluated throw TF (20) taking into account that $H^{mn}(k\omega_n^{inst}) = H^{mn}(k\dot{\theta}_n)$. The replacement of the constant factor K^{mn} by the frequency-dependent TF $H^{mn}(k\dot{\theta}_n)$ leads to the following form of KM.

$$\frac{d\theta_m}{dt} = \omega_m + \sum_{k=-K}^{K} u_{mn}\left(\frac{d\theta_n}{dt}, \theta_m - \theta_n\right). \quad (22)$$

Here coupling function $u_{mn}(\dot{\theta}_n, \Delta\theta)$ is 2π-periodic with respect to $\Delta\theta$ and its Fourier magnitudes represent generalization of (19):

$$U_k^{mn}\big(\dot{\theta}_n\big) = H^{mn}\big(\dot{\theta}_n\big)\omega_m V_k^m X_k^n. \quad (23)$$

Equation (22) is ODE system unresolved with respect to derivatives. Note that in contrast to the inclusion of ODE equations (5) into the PPV macromodel using $H^{mn}\big(\dot{\theta}_n\big)$ obtained by solving (20) does not increase the size of the ODE system.

IV. Kuramoto Model with a Fractional Ratio of Oscillators' Frequencies

Equations (22) were derived assuming close values of the oscillators' fundamentals ω_m . Here we analyze oscillators with ratios of fundamentals close to the rational fraction defined by coprimes p_m, p_n

$$\frac{\omega_m}{\omega_n} \approx \frac{p_m}{p_n} or \frac{\omega_m}{p_m} \approx \frac{\omega_n}{p_n}. \quad (24)$$

Firstly, we consider the simple case (2), (18) with

$$\omega_0/\omega_{ex} \approx p/q \; . \quad (25)$$

To solve the problem, one can take into account that periodic oscillation with period T (frequency f=1/T) and Fourier harmonics A_i (i=0,…,I) is also represented as the oscillation with any multiple period $\overline{T} = mT$ ($\overline{f} = 1/\overline{T}$) and harmonics $\overline{A}_j$ (j=0,…,mI). These harmonics have the values:

$$\overline{A}_{m\cdot i} = A_i, \; \overline{A}_j = 0 \; for \; j \neq m \cdot i. \quad (26)$$

In other words, the sequence of $\overline{A}_j$ is formed by inserting m-1 zeroes after each value in the sequence of A_i.

Thus, for two oscillators (25) we can introduce new fundamentals and instantaneous phases as

$$\overline{\omega}_0 = \frac{\omega_0}{p} \approx \; \overline{\omega}_{ex} = \frac{\omega_{ex}}{q}, \overline{\theta} = \frac{\theta}{p}, \overline{\theta}_{ex} = \frac{\theta_{ex}}{q} \; . \quad (27)$$

Then (2) is presented as $d\overline{\theta}/dt = \overline{\omega}_0 + u_{ex}(\overline{\theta} - \overline{\theta}_{ex})$. After replacing variables by (27) we obtain

$$\frac{d\theta}{dt} = \omega_0 + pu_{ex}\left(\frac{\theta}{p} - \frac{\theta_{ex}}{q}\right). \quad (28)$$

In the expression for Fourier series of $u_{ex}\big(\Delta\overline{\theta}\big)$ transformed by (27) harmonics X_k^{ex}, V_k are replaced by $\overline{X_k^{ex}}, \overline{V}_k$. Then nonzero term can appear if both $\overline{X_k^{ex}}, \overline{V}_k$ are nonzero. Due to (26) it occurs for indexes $k = l \cdot p \cdot q$ for an integer l. Thus (18) can be transformed to the expression without superfluous zeros:

$$u_{ex}\big(\Delta\overline{\theta}\big) \; = \sum_{l=-L}^{L} K^{ex}\overline{\omega}_0 X_{lpq}^{ex} V_{lpq} exp\big(jlpq\Delta\overline{\theta}\big) \; .$$

Taking into account that $p \cdot q \cdot \Delta\overline{\theta} = q\theta - p\theta_{ex}$ and replacing numeration index l by k we obtain

$$pu_{ex}\big(\Delta\overline{\theta}\big) = \overline{u}_{ex}(q\theta - p\theta_{ex}) =$$

$$= K^{ex}\omega_0 \sum_{l=-L}^{L} X_{kpq}^{ex} V_{kpq} \; exp \, jk(q\theta - p\theta_{ex}). \quad (29)$$

Then KM presented in (28) has the form in source variables:

$$\frac{d\theta}{dt} = \omega_0 + \overline{u}_{ex}(q\theta - p\theta_{ex}), \quad (30)$$

where function $\overline{u}_{ex}$ is defined by (29).

To expand the obtained results for excited oscillator onto the oscillatory system (1) we firstly introduce generic form of the

coupling function as $U(X,V,H,p/q\,,\omega_{inst},\delta)$. Here X is the waveform vector of the exciting oscillator, V is the PPV vector of the perturbed oscillator, H is the transfer function of the interconnect between the oscillators, ω_{inst} is the instantaneous frequency of the exciting oscillator, δ is the discrepancy between instantaneous phases of the oscillators (e.g. in (30) $\delta = q\theta - p\theta_{ex}$). Similarly with presented above derivation of (29) one can obtain from (23) Fourier expansion for the function $U(X,V,H,p/q\,,\delta)$:

$$U(X,V,H,p/q\,,\omega_{inst},\delta) = \\ = \sum_{l=-L}^{L} H(\omega_{inst})X_{kpq}V_{kpq}\,exp\,jk(\delta). \quad (31)$$

To apply (31) one needs to define rational fractions p_{ij}/q_{ij} corresponding to the frequency ratios of each pair of coupled oscillators (24). It can be performed if frequency ratios of the first oscillator to any *i*-th oscillator are given close to fractions

$$\omega_1/\omega_i \approx p_i/q_i. \quad (32)$$

One can obtain from (32) the rational frequency ratios for any two oscillators satisfying equalities:

$$\frac{\omega_i}{\omega_j} = \frac{\omega_1/\omega_j}{\omega_1/\omega_i} \approx \frac{p_j/q_j}{p_i/q_i} = \frac{p_j q_i}{p_i q_j} = \frac{p_{ij}}{q_{ij}}. \quad (33)$$

To provide p_{ij}, q_{ij} be coprime the following operations can be performed:

$$r = gcd\big(p_j q_i, p_i q_j\big), p_{ij} = p_j q_i/\mathrm{r}\,, q_{ij} = p_i q_j/\mathrm{r}. \quad (34)$$

Here $gcd(a,b)$ denotes the greatest common divisor of a,b.

After executing (33), (34) p_{ij}, q_{ij} are known and we can parametrize the coupling functions applying generic form (31)

$$u_{mn}(\,p/q\,,\omega_{inst},\delta) = \\ = \omega_m U(X^n, V^m, H^{mn},\, p_{mn}/q_{mn}\,,\omega_{inst},\delta). \quad (35)$$

The parametrized coupling functions (35) fully determine KM ODE system. This system is finally presented as

$$\frac{d\theta_m}{dt} = \omega_m + \sum_{n=1}^{N} u_{mn}\left(p_{mn}\frac{d\theta_n}{dt}\,, q_{mn}\theta_m - p_{mn}\theta_n\right). (36)$$

Note that equation (36) coincides with (22) in the particular case of close frequencies: $p_{mn} = q_{mn} = 1$.

V. CONCLUSIONS

The paper proposes modifications of the Kuramoto model providing its application to problems of the circuit simulation.

Proposed parametrized Kuramoto model is based on harmonics of periodic steady state solution and harmonics of perturbation projection vectors that can be obtained by the simulation of each oscillator circuit in free-running mode.

Dynamic couplings defined by frequency-depended transfer functions of linear interconnects are included into the model. The frequency argument of the transfer function is defined by the time derivative of the corresponding phase variable. Such an approach does not increase the order of the model's ODE system.

The problem of analyzing a system of coupled oscillators with natural frequency ratios close to rational fractions is considered for the first time. The new approach is based on bringing the oscillator frequencies to a common value by converting the frequency of each oscillation using a multiple period.

REFERENCES

[1] P. Maffezzoni, B. Bahr, Z. Zhang and L. Daniel, "Oscillator Array Models for Associative Memory and Pattern Recognition," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 62, no. 6, pp. 1591-1598, June 2015. doi: 10.1109/TCSI.2015.2418851

[2] P. Maffezzoni, B. Bahr, Z. Zhang and L. Daniel, "Reducing Phase Noise in Multi-Phase Oscillators, "IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 63, no. 3, pp. 379-388, March 2016.

[3] J. A. Acebrón, et al. "The Kuramoto model: A simple paradigm for synchronization phenomena," Reviews of Modern Physics 77(1), 137 – 185 (2005).

[4] M. Bonnin, F. Corinto and M. Gilli, "Periodic Oscillations in Weakly Connected Cellular Nonlinear Networks, "IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 55, no. 6, pp. 1671-1684, July 2008. doi: 10.1109/TCSI.2008.916460

[5] P. Ashwin, S. Coombes, R. Nicks, "Mathematical Frameworks for Oscillatory Network Dynamics in Neuroscience, "*J. Math. Neurosc.* **6,** 2 (2016) doi:10.1186/s13408-015-0033-6

[6] D.J. Jörg, et al., "Synchronization Dynamics in the Presence of Coupling Delays and Phase Shifts," Phys Rev Lett. 112(17): 174101, (2014).

[7] C. A. S. Batista, et al., "Synchronization of phase oscillators with coupling mediated by a diffusing substance," Physica A, 470, 236-248 (2017).

[8] J.M.V. Grzybowski, et al. "On synchronization in power-grids modelled as networks of second-order Kuramoto oscillators," Chaos 26(11) , 113113 (2016)

[9] J. M. T. Thompson and H. B. Stewart, Nonlinear Dynamics and Chaos, Wiley, 1986.

[10] A. Demir, A. Mehrotra, J. Roychowdhury," Phase Noise in Oscillators: A Unifying Theory and Numerical Methods for Characterization," IEEE Trans. on CAS I. 47(5), 655, (2000)

[11] A. Demir, J. Roychowdhury. "A reliable and efficient procedure for oscillator PPV computation with phase noise macromodeling applications," IEEE Trans. on Comp.-Aided Design of Integrated Circuits and Systems 22(2), 188-197 (2003).

[12] D. Harutyunyan, et al.: "Simulation of mutually coupled oscillators using nonlinear phase macromodels and model order reduction techniques," In: M. Gunter (Ed.): Coupled Multiscale Simulation and Optimization in Nanoelectronics, Mathematics in Industry 21, pp. 398-432. Springer, Berlin, Heidelberg (2015).

[13] M.M. Gourary, S.G. Rusakov, et al. "Smoothed Form of Nonlinear Phase Macromodel for Oscillators," In: IEEE/ACM Int. Conf. on Comp.-Aided Design, pp. 807 - 814 (2008).

[14] M.M. Gourary, S.G. Rusakov, et al. "Injection Locking Conditions under Small Periodic Excitations." In: 2008 IEEE Int. Symp. on Circ. and Syst., pp. 544-547 (2008).

Ternary Sum Codes

Dmitry Efanov,
DSc, Professor at Higher School of Transport,
Institute of Mechanical Engineering, Materials and Transport,
Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation
TrES-4b@yandex.ru

Abstract—The paper describes the research results in the field of ternary codes construction focused on the use of checkable digital devices and systems and their diagnostic support in the synthesis tasks. The article provides a method for constructing a ternary sum code, which is analogous to the classical binary Berger code. The article identifies the previously unknown properties of error detection by the ternary sum code in the case of their occurrence only in the data vectors with the faultlessness of check bits. This task is relevant in practical applications where the bits of the check and data vectors are calculated physically by different devices. In addition, a comparison of binary and ternary sum codes is carried out. The article shows that the share of undetectable errors with the d multiplicity from the total number of errors with this multiplicity in the sum codes is a constant value and does not depend on the length of the data vector. This property is inherent in both binary and ternary sum codes.

Keywords—devices and systems operating in ternary logic; ternary sum code; Berger code; undetectable error; code properties.

I. Introduction

In the modern world, the synthesis of digital devices and systems uses the binary logic. Nevertheless, all over the world, research has been and is being conducted in the field of the using and the construction of devices and systems that function in ternary logic [1 – 9]. First of all, the advantage of the ternary logic is a denser recording of numbers and the possibility of using it. The devices functioning in the ternary logic were built in the second half of the last century [2]. There are also the modern implementations of devices that operate on the similar principles [3]. Some researchers note that using the ternary logic instead of the binary has advantages in the quantum computers implementing [10]. These examples demonstrate the relevance of research in the field of the ternary logic using in the construction of digital devices and systems.

It is widely known that the methods of redundant coding, backing up, diversification, etc. of both hardware and software are used in the construction of the modern reliable and safety devices and automation systems [11 – 14]. All these methods are applied at various levels of device and system architecture (both at the micro level in using a highly reliable element base, and at the macro level while ensuring the reliability of systems functioning during data transfer between components and control objects). The redundant coding is also used in the synthesis of checkable digital devices and systems and their diagnostic support [15 – 18]. Such applications often use the codes that are oriented to the error detection rather than correction of this. Firstly, the error-correcting codes have a little more redundancy than the error-detecting codes. Secondly, the correction of errors in the hardware can lead to their accumulation and subsequent critical failure of the system. The constant-weight codes, various sum codes and the codes with parity checks of bit values are used everywhere [19 – 25].

The paper is devoted to the presentation of research results in the field of construction of ternary codes focused on the use of checkable digital devices and systems and their diagnostic support in the synthesis tasks, as well as the research of their characteristics. The article describes the ternary sum codes and some of their properties, which should be taken into account in the synthesis of reliable and safety devices and systems operating in the ternary logic.

II. The Principles of the Sum Code Construction

There are a wide variety of the ternary error-tolerant codes constructing methods [26 – 30]. There are several ways for the code construction. The first way is to create the heuristic rules for the check bits obtaining. The second way is to identify the primary basic properties of the "future" code. Let's construct a ternary code that will detect any errors related to violation of the number of bits equal to 1 and 2 in the data vectors. Such errors belong to the monotonous and asymmetric classes, and the constructed code belongs to the *MAED*-codes class (monotonous & asymmetric error-detection codes). These codes can be used in the synthesis of the checkable structures of the devices operating in the ternary logic, as well as their diagnostic support.

The ternary code belonging to the class of codes with the any monotonous and asymmetric error detection (to the class of *MAED*-codes) is constructed as follows.

Algorithm 1. *The rules for determining the values of bits for the check vectors of ternary sum codes:*

1. In the data vector with the length m, it is necessary to determine the number of bits equal to 1 and the number of bits equal to 2 (these are the numbers r_1 and r_2).
2. The number r_1 is represented in the ternary form and is fixed in the $k_1 = \lceil \log_3(m+1) \rceil$ upper bits of the check vector (the notation $\lceil ... \rceil$ denotes an integer above the calculated value).
3. The number r_2 is represented in the ternary form and is fixed in the $k_2 = \lceil \log_3(m+1) \rceil$ lower bits of the check vector.

Because the numbers r_1 and r_2 are determined by using summation operations, we call the code constructed by algorithm 1 a sum code and denote it as $\Sigma(m,k)$-code, where $k = k_1 + k_2 = 2\lceil \log_3(m+1) \rceil$ is the number of bits in the check vectors. The Table 1 shows the values of the number k for $\Sigma(m,k)$-codes with the different lengths of the data vec-

978-1-7281-9899-6/20/$31.00 ©2020

tor. The table shows that Σ(*m*,*k*)-codes are acceptable to construct for the cases *m*>4 (otherwise, it is possible to use the ternary codes with repetition).

TABLE I. THE NUMBER OF CHECK BITS IN Σ(M,K)-CODES

m	*k*
4	4
5	4
6	4
7	4
8	4
9	6
10	6
…	…
20	6
…	…
50	8
…	…
100	10

We give an example of obtaining a check vector for the data vector <01012121222>. The length of the data vector is *m*=11. It follows that the numbers $k_1 = k_2 = \lceil \log_3 12 \rceil = 3$, and k=6. The number r_1=4 corresponds to the ternary vector <011>. The number r_2=5 corresponds to the ternary vector <012>. Thus, the check vector of the Σ(12,6)- code will have the form <011012>.

III. SOME PROPERTIES OF TERNARY SUM CODES

Let's consider the features of error detection in data vectors by Σ(*m*,*k*)-codes.

First of all, we note that Σ(*m*,*k*)-codes, with the exception of codes with data vectors with the length $m = 3^p - 1$, $p \in \{2,3,4,...\}$, do not use all possible combinations of bits in the check vectors. The Σ(*m*,*k*)-codes with the lengths of data vectors $m = 3^p - 1$, $p \in \{2,3,4,...\}$, are called *the perfect ternary sum codes*.

Proposition 1. *Σ(m,k)-codes detect any monotonous and asymmetrical errors in data vectors and do not detect any compositional errors.*

The validity of the Proposition 1 follows from the principles of the Σ(*m*,*k*)-code construction. In fact, in the code construction the data vectors are classified into check groups corresponding to numbers r_1 and r_2 (r_1-r_2 groups), which determine all data vectors with the same composition. An error will be undetectable if it distorts the data vector belonging to one check group into the data vector belonging to the same check group. This error does not disturb the composition of values and is compositional. It follows that the compositional errors cannot be detected by Σ(*m*,*k*)-codes. If the error transfers the data vector of one check group to the data vector of another check group, it will be associated with a violation of the composition of values in the data vector. This error will be detected and will belong to the type of monotonous or asymmetrical errors.

The table form of the code specification, in which all data vectors are classified into all possible r_1-r_2 groups, should be used to determine the characteristics of error detection by Σ(*m*,*k*)-codes. It is possible to determine the features of error detection by Σ(*m*,*k*)-codes by analyzing such r_1-r_2 groups. For example, Table 2 defines a Σ(4,4)-code.

The Σ(4,4)-code specification table includes data vectors divided into three categories. The first category includes only those data vectors for which the bits take only the values 0 and 1. Such data vectors occupy r_1-0 groups. The second category of data vectors includes the vectors for which the bits take only the values 0 and 2 and occupy 0-r_2 groups. The third category includes data vectors, the bits in which take values 0, 1 and 2. This division of data vectors into groups and categories makes it possible to determine the main characteristics of error detection by Σ(*m*,*k*)-codes.

Note also that it is possible to separately consider sum codes, for which only data vectors are used that belong to one of three categories: $\Sigma^1(m,k)$, $\Sigma^2(m,k)$, $\Sigma^3(m,k)$ codes. If, for example, we consider $\Sigma^1(m,k)$-code corresponding to the distribution of the data vectors only of the first category, we can say that the classical binary sum code (Berger code) is constructed [31].

Proposition 2. *Undetectable errors that occur in data vectors that belong to the r_1-0 and 0-r_2 groups can only have an even multiplicity.*

The statement of Proposition 2 is supported by the following considerations. Each r_1-r_2 group for the cases under consideration is defined only by the number r_1 (or only by the number r_2). In order for the error to be undetectable, it must transfer the data vectors of a particular check group into each other. In this case, the total number of bits equal to 1 (or equal to 2) must not be violated. This is only possible if any distortion 1 (or 2) is compensated by the opposite zero-bit distortion. It follows that the error will have only an even multiplicity.

Undetectable errors for the check groups characterizing data vectors of the third category can have any multiplicity.

The Table 2 follows a method for calculating the number of errors undetectable by Σ(*m*,*k*)-codes: for a given value *m*, it is necessary to determine the total number of the data vectors corresponding to each r_1-r_2 group. Let's denote this number as Q_{r_0,r_1,r_2}, where r_0, r_1, r_2 are the numbers of bits equal to 0, 1, and 2, respectively. The Table 3 gives the representatives of the check groups of the Σ(4,4)-code. The number of data vectors corresponding to each check group can be determined by the formula:

$$Q_{r_0,r_1,r_2} = C_m^{r_1} C_{m-r_1}^{r_2} C_{m-(r_1+r_2)}^{r_0}, \quad (1)$$

where $C_m^{r_1}$ is the number of variants of the location of bits equal to 1 in the data vector of the length *m*; $C_{m-r_1}^{r_2}$ is the number of variants of the location of bits equal to 2 in the remaining $m - r_1$ data vectors; $C_{m-(r_1+r_2)}^{r_0}$ is the number of variants of the location of bits equal to 0 in the remaining $m - (r_1 + r_2)$ data bits.

Taking into account that $m - (r_1 + r_2) = r_0$ and $C_{r_0}^{r_0} = 1$, we rewrite formula (1) in the form:

$$Q_{r_0,r_1,r_2} = C_m^{r_1} C_{m-r_1}^{r_2}. \quad (2)$$

TABLE II. THE DISTRIBUTION OF THE DATA VECTORS INTO CHECK GROUPS OF THE Σ(4,4)-CODE

Check groups (r_1-r_2 groups)														
00-00	00-01	00-02	00-10	00-11	01-00	01-01	01-02	01-10	02-00	02-01	02-02	10-00	10-01	11-00
Data vectors														
0000					0001 0010 0100 1000				0011 0101 0110 1001 1010 1100			0111 1011 1101 1110		1111
	0002 0020 0200 2000	0022 0202 0220 2002 2020 2200	0222 2022 2202 2220	2222										
						0012 0021 0102 0120 0201 0210 1002 1020 1200 2001 2010 2100	0122 0212 0221 1022 1202 1220 2012 2021 2102 2120 2201 2210	1222 2122 2212 2221		0112 0121 0211 1012 1021 1102 1120 1201 1210 2011 2101 2110	1122 1212 1221 2112 2121 2211		1112 1121 1211 2111	

TABLE III. THE REPRESENTATIVES OF THE CHECK GROUPS OF THE Σ(4,4)-CODE

r_1-r_2 group	The representatives of the check groups	The total number of representatives	The formula for calculation
00-00	0000	1	$C_4^0 C_4^0$
00-01	0002	4	$C_4^0 C_4^1$
00-02	0022	6	$C_4^0 C_4^2$
00-10	0222	4	$C_4^0 C_4^3$
00-11	2222	1	$C_4^0 C_4^4$
01-00	0001	4	$C_4^1 C_3^0$
01-01	0012	12	$C_4^1 C_3^1$
01-02	0122	12	$C_4^1 C_3^2$
01-10	1222	4	$C_4^1 C_3^3$
02-00	0011	6	$C_4^2 C_2^0$
02-01	0112	12	$C_4^2 C_2^1$
02-02	1122	6	$C_4^2 C_2^2$
10-00	0111	4	$C_4^3 C_1^0$
10-01	1112	4	$C_4^3 C_1^1$
11-00	1111	1	$C_4^4 C_0^0$

Then the number of undetectable errors in every check group will be determined by the value:

$$N_{r_0,r_1,r_2} = C_m^{r_1} C_{m-r_1}^{r_2} \left(C_m^{r_1} C_{m-r_1}^{r_2} - 1 \right). \quad (3)$$

The total number of undetectable errors is equal to the sum of all undetectable errors "given" by each check group:

$$N_{m,k} = \sum_{r_1,r_2=0}^{r_1,r_2=m} N_{r_0,r_1,r_2} = \sum_{r_1=0}^{r_1=m} \left(\sum_{r_2=0}^{r_2=m} C_m^{r_1} C_{m-r_1}^{r_2} \left(C_m^{r_1} C_{m-r_1}^{r_2} - 1 \right) \right). \quad (4)$$

For example, using formula (4) to calculate the total number of errors undetectable by $\Sigma(m,k)$-code gives the following result: $N_{4,0,0} = 0$, $N_{3,0,1} = 12$, $N_{2,0,2} = 30$, $N_{1,0,3} = 12$, $N_{0,0,4} = 0$, $N_{3,1,0} = 12$, $N_{2,1,1} = 132$, $N_{1,1,2} = 132$, $N_{0,1,3} = 12$, $N_{2,2,0} = 30$, $N_{1,2,1} = 132$, $N_{0,2,2} = 30$, $N_{1,3,0} = 12$, $N_{0,3,1} = 12$, $N_{0,4,0} = 0$. Summing up the obtained numbers, we get: $N_{4,4} = 558$.

When the length of the data vector increases by one, the number of the check groups (the representatives of the check groups) also increases. Moreover, this pattern is captured: in the 00-00 group there is always 1 vector, in the 00-$r_2 - m$, 01-$r_2 - m$, 02-$r_2 - m$–1, 11-$r_2 - m$–2, …, r_1-00 groups there is always 1 vector. Based on this, the total number of the check groups (and various compositions) is determined by the formula:

$$\begin{aligned} R_m &= 1 + m + m + (m-1) + (m-2) + \ldots + 2 + 1 = \\ &= m + 1 + \frac{m(m+1)}{2} = \frac{m^2 + 3m + 2}{2}. \end{aligned} \quad (5)$$

For example, for the considered $\Sigma(4,4)$-code $R_4 = \frac{4^2 + 3 \cdot 4 + 2}{2} = 15$.

The Table 4 shows the calculated values of the R_m number for various $\Sigma(m,k)$-codes.

TABLE IV. THE R_M NUMBERS FOR DIFFERENT $\Sigma(M,K)$-CODES

m	R_m
4	15
5	21
6	28
7	36
8	45
9	55
10	66
…	…
20	231
…	…
50	1326
…	…
100	5151

Let's start considering the relative error detection indicators of the $\Sigma(m,k)$-codes.

The total number of errors that can occur in the data vectors of ternary codes is equal to [32]:

$$N_m = 3^m \left(3^m - 1 \right). \quad (6)$$

The $\gamma_{m,k}$ indicator makes it possible to compare the number of errors undetectable by codes with the total possible number of errors:

$$\gamma_{m,k} = \frac{N_{m,k}}{N_m} \cdot 100\%. \quad (7)$$

The value of $\gamma_{m,k}$ is closer to zero, the greater the number of errors detected by the considered code.

As shown in [32], there is such a ternary separable code that has a uniform distribution of data vectors into check groups. This code has a very important feature.

Theorem 1. *A ternary code with the m and k parameters will have the minimum total number of undetectable errors, when the following conditions are met: all 3^m data vectors are distributed evenly among all 3^k check vectors, and the total number of undetectable errors in such code will be determined by the formula:*

$$N_{m,k}^{\min} = 3^m \left(3^{m-k} - 1 \right). \quad (8)$$

Thus, any separable code, including the $\Sigma(m,k)$-code, can be compared with a code with a uniform distribution of data vectors into check groups:

$$\xi_{m,k} = \frac{N_{m,k}}{N_{m,k}^{\min}} \cdot 100\%. \quad (9)$$

The indicator $\xi_{m,k}$ characterizes the efficiency of the check bits using by the ternary separable code: the closer it is to one, the more effectively the code detects errors.

The Table 5 shows the characteristics of $\Sigma(m,k)$-codes with small values of the data vector lengths calculated using the above formulas. The graphs of the dependences of the $\gamma_{m,k}$ and $\xi_{m,k}$ on m values complement the above calculated numbers.

As the value of m increases, the share of undetectable errors from their total number decreases. The efficiency of the check bits using increases when the length of the data vector increases to the value $m = 3^p - 1$, $p \in \{2,3,4,...\}$, corresponding to the value of the length of the data vector of the perfect $\Sigma(m,k)$-code. Upon reaching the value $m = 3^p$, $p \in \{2,3,4,...\}$, there is a sharp decrease in the $\xi_{m,k}$ coefficient, which is explained by an increase in the number of check bits by 2 and an extremely large number of unused check groups. Then, gradually, as the length of the data vector increases, the number of data vectors grows, a larger number of check groups are filled, and the $\xi_{m,k}$ coefficient increases.

$\Sigma(m,k)$-codes do not detect less than 10% of errors in data vectors for any length. Moreover, at $m \geq 8$, the share of undetectable errors from their total number is less than 5%. However, the check bits are used very inefficiently, as shown by the graph of the $\xi_{m,k}$ value: the values of this coefficient are no more than 25% for codes with any values of m.

In comparison with the binary sum codes ($\Sigma^1(m,k)$-codes), the ternary sum codes have significantly lower values of the $\gamma_{m,k}$ and $\xi_{m,k}$ indicators at the same values of m (see the Table 6). In this case, $\Sigma^1(m,k)$- codes detect much more errors than $\Sigma(m,k)$-codes. The last column of the Table 6 shows the values of θ_m, which shows how many times more errors are detected by $\Sigma^1(m,k)$-codes compared to $\Sigma(m,k)$-codes. This case, obviously, follows from the fact that there is much smaller number of errors in the code vectors of binary logic than in the code vectors of ternary logic. The data in the Table 6 are supplemented by the graphs of changes of the $\gamma_{m,k}$ and $\xi_{m,k}$ values with an increase of m, which are shown in Fig. 1 and Fig. 2.

The above formulas, unfortunately, do not make it possible to calculate the number of errors undetectable by $\Sigma(m,k)$-codes for every multiplicity. This requires a detailed analysis of the tables defining the codes. The results of this analysis for the $\Sigma(m,k)$-codes with small values of the data vector lengths are presented in the Table 7. The table shows the following three numbers for each m value. The first number (it is written in the top line) characterizes the total number of undetectable errors with the d multiplicity. The second number (it is written in the middle line) is the total number of errors of this multiplicity. The third number (it is written in the bottom line) is the share of undetectable errors with the d multiplicity (the β_d, % value).

Calculations show that $\Sigma(m,k)$-codes (as well as their binary analogues [33]) have an interesting feature.

Theorem 2. *The share of errors with the d multiplicity from the total number of errors with this multiplicity that are undetectable by ternary sum codes does not depend on the length of the data vector and is a constant value.*

Thus, any $\Sigma(m,k)$-codes don't detect 16.667 % twofold errors, 5.556 % triple errors, and 6.944 % quadruple errors, etc. However, it should be noted that the statement of the Theorem 2 is a hypothesis, which is confirmed by calculations, but has not yet been mathematically confirmed.

TABLE V. THE INDICATORS OF ERROR-DETECTION BY $\Sigma(M,K)$-CODES

m	k	$N_{m.k}$	N_m	$N_{m,k}^{\min}$	$\gamma_{m.k}$	$\xi_{m.k}$
4	4	558	6480	0	8.611	0
5	4	4410	58806	486	7.499	11.02
6	4	34440	530712	5832	6.489	16.934
7	4	270648	4780782	56862	5.661	21.01
8	4	2151198	43040160	524880	4.998	24.399
9	6	17300154	387400806	511758	4.466	2.958
10	6	140609016	3486725352	4723920	4.033	3.36
11	6	1153285848	31380882462	42869574	3.675	3.717
12	6	9533107584	282429005040	386889048	3.375	4.058
13	6	79324972272	2541864234006	3485190078	3.121	4.394
14	6	663830247366	22876787671992	31376276640	2.902	4.727
15	6	5582710119186	205891117745742	282415187574	2.711	5.059

TABLE VI. THE CHARACTERISTICS OF ERROR-DETECTION BY $\Sigma(M,K)$- AND $\Sigma^1(M,K)$-CODES

m	$\gamma_{m.k}$		$\xi_{m.k}$		θ_m
	$\Sigma(m.k)$	$\Sigma^1(m.k)$	$\Sigma(m.k)$	$\Sigma^1(m.k)$	
4	8.611	48.214	0	29.63	10.333
5	7.499	45.833	11.02	43.636	20.045
6	6.489	43.347	16.934	52.093	40.047
7	5.661	40.972	21.01	58.111	81.915
8	4.998	38.798	24.399	30.442	170.541
9	4.466	36.847	2.958	32.992	359.611
10	4.033	35.113	3.36	35.112	765.294
11	3.675	33.573	3.717	36.978	1639.625
12	3.375	32.203	4.058	38.684	3530.702
13	3.121	30.979	4.394	40.28	7632.973
14	2.902	29.881	4.727	41.797	16554.281
15	2.711	28.889	5.059	43.251	35997.802

The Table 8 shows for comparison the values of β_d indicators for the binary and ternary sum codes ($\Sigma(m,k)$- and $\Sigma^1(m,k)$-codes). The η_d value given in the last column shows how many times the β_d value for the $\Sigma(m,k)$-code is smaller than the same value for the $\Sigma^1(m,k)$-code. The binary sum codes detect any errors of odd multiplicity, and the ternary sum codes do not have this feature. The β_d values for ternary sum codes are significantly lower than for binary sum codes (for example, the binary sum codes do not detect 50 % of twofold errors [33], and ternary codes do not detect 16.667 % errors). This feature of ternary sum codes is related to the principles of their construction and a much larger total number of errors than in code vectors of binary logic. In addition, it is also possible to note the different nature of the change in the dependence of β_d on the value of d as it increases. For the $\Sigma^1(m,k)$-code, this indicator gradually decreases with increasing of d (if only even multiplicities are considered). For the $\Sigma(m,k)$-code, there is also a decrease in the value of β_d with an increase of d, but it is not monotonic.

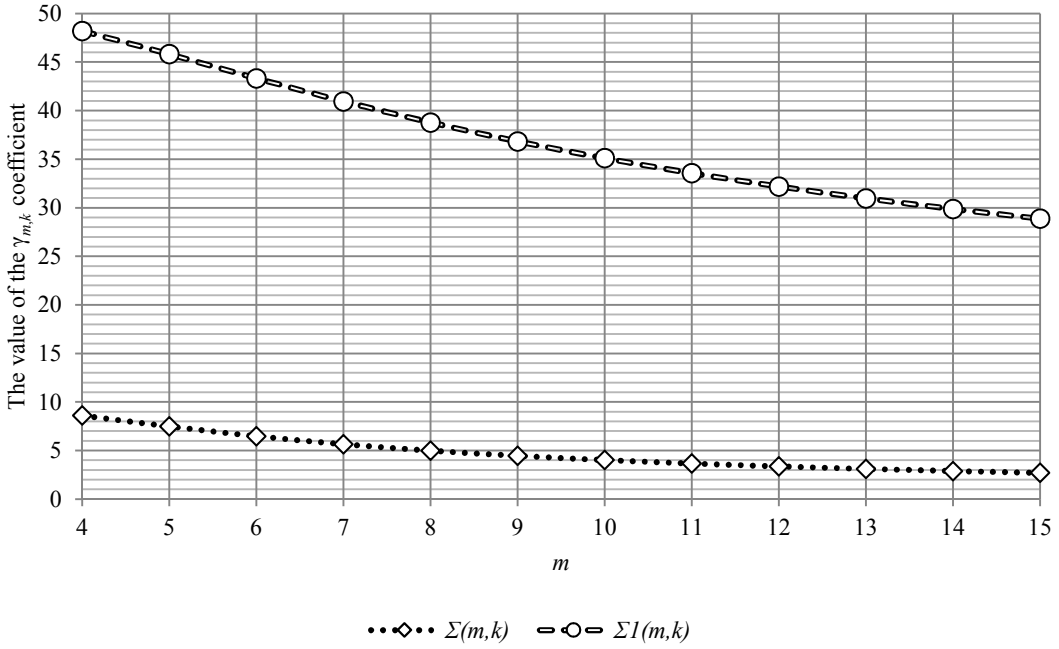


Fig. 1. The dependencies of the $\gamma_{m,k}$ coefficients on the value of m for $\Sigma(m,k)$- and $\Sigma^1(m,k)$-codes.

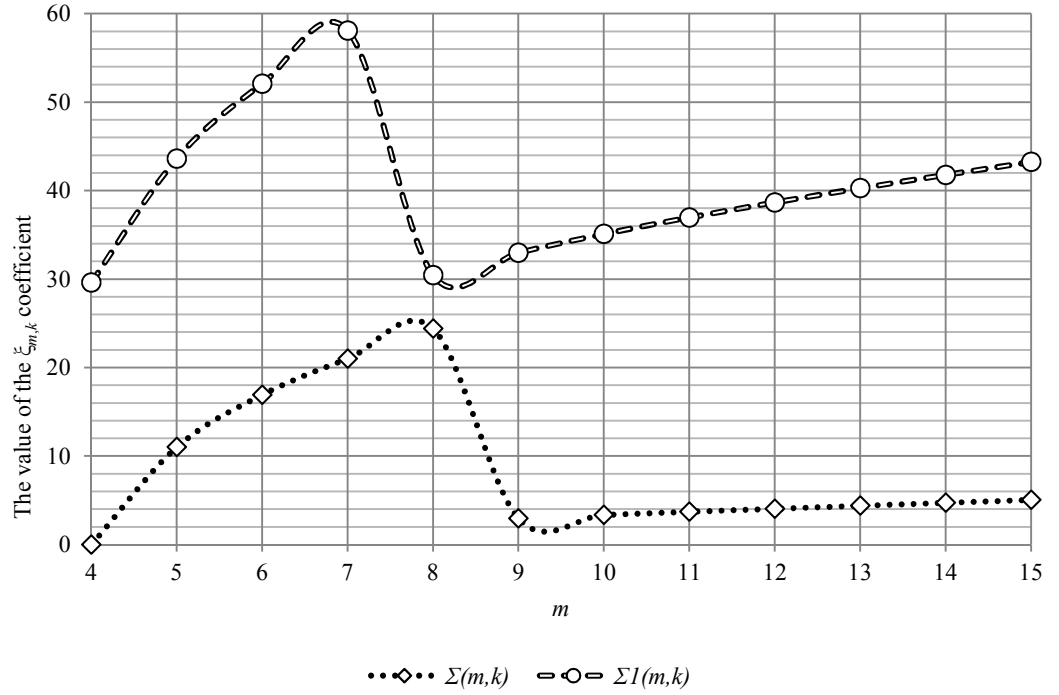


Fig. 2. The dependencies of the $\xi_{m,k}$ coefficients on the value of m for $\Sigma(m,k)$- and $\Sigma^1(m,k)$-codes.

TABLE VII. THE CHARACTERISTICS OF ERROR DETECTION BY $\Sigma(M,K)$-CODES BY MULTIPLICITY

m	d					
	1	2	3	4	5	6
3	0	54	12	–	–	–
	162	324	216	–	–	–
	0	16.667	5.556	–	–	–
4	0	324	144	90	–	–
	648	1944	2592	1296	–	–
	0	16.667	5.556	6.944	–	–
5	0	1620	1080	1350	360	–
	2430	9720	19440	19440	7776	–
	0	16.667	5.556	6.944	4.63	–
6	0	7290	6480	12150	6480	2040
	8748	43740	116640	174960	139968	46656
	0	16.667	5.556	6.944	4.63	4.372

TABLE VIII. THE COMPARISON OF $\Sigma(M,K)$- AND $\Sigma^1(M,K)$-CODES BY B_D INDICATOR

d	$\Sigma(m,k)$	$\Sigma^1(m,k)$	η_d
2	16.667	50	3
3	5.556	0	0
4	6.944	37.5	5.4
5	4.63	0	0
6	4.372	31.25	7.148

IV. CONCLUSION

The method of the ternary code construction presented in the article makes it possible to obtain a code that will have the property of any monotonous and asymmetrical error-detection in data vectors. However, at the same time, this code will not detect any errors of the composite type. The share of such errors for all $\Sigma(m,k)$-codes with $m \geq 8$ is less than 5% from the total number of errors in data vectors. This is not so significant compared to the binary sum codes (Berger codes). When comparing the ternary sum code with Berger codes [33], it is possible to note a much higher share of the errors detected by the ternary sum code. Due to the noted features, the ternary sum codes can be widely used in the synthesis of digital devices and systems operating in the ternary logic.

Let's focus on a certain feature of the $\Sigma(m,k)$-codes. All possible check vectors are used only in special cases – for each value of k there is only one perfect sum code, for which $m = 3^p - 1, \quad p \in \{2,3,4,...\}$. All other sum codes use the check vectors very inefficiently (see, for example, the Tables 2 and 5). The presented feature of the $\Sigma(m,k)$-codes complicates the synthesis of fully self-checking structures of their coders.

The ternary sum codes presented in this paper can be quite simply modified. The following method can be used for the code construction. The numbers r_1 and r_2 are calculated in the $M \in \{3^1, 3^2, ..., 3^{\lceil \log_3(m+1) \rceil - 1}\}$ modulus residue ring. The obtained values are recorded in the bits of the ternary check vector. For example, one of these codes is the code for which the numbers r_1 and r_2 are calculated in the M=9 modulus residue ring. The number of check bits in such ternary modular codes will always be equal to k=4. This code will use the bits of check vectors more effectively, and also this code will detect any monotonous error in the data vectors of the multiplicity d_μ<9. For this reason their application may also be perspectival.

In conclusion, we note that the presented method for ternary code construction is one of many. The further research can be directed to identification of the features of the various modifications of the $\Sigma(m,k)$-code, as well as to identification of the main characteristics of their detection of errors in data vectors. The results of these studies in the future may form the basis for the construction of self-checking and checkable digital devices and systems operating in ternary logic.

REFERENCES

[1] 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.

[2] D. Roy, and Jr. Merril "Ternary Logic in Digital Computers", Proc. of DAC '65, ACM New York, NY, USA, pp. 6.1-6.17, doi: 10.1145/800266.810759.

[3] J. Connely "Ternary Computing Testbed 3-Trit Computer Architecture", California Polytechnic State University of San Luis Obispo, August 29th, 2008, 184 p.

[4] C. Vudadha, S. Katragadda, and P.S. Phaneendra "2:1 Multiplexer Based Design for Ternary Logic Circuits", IEEE Asia Pacific Conf. on Postgraduate Research in Microelectronics and Electronics (PrimeAsia), 19-21 December 2013, Visakhapatnam, India, pp. 46-51, doi: 10.1109/PrimeAsia.2013.6731176.

[5] R.S.P. Nair, S.C. Smith, and J. Di "Delay Insensitive Ternary CMOS Logic for Secure Hardware", J. of Low Power Electronics and Applications, 2015, Issue 5, pp. 183-215, doi:10.3390/jlpea5030183.

[6] Md.R. Rahman, and J.E. Rise "On Designing a Ternary Reversible Circuit for Online Testability", IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing, 2011, pp. 1-7, doi: 10.1109/PACRIM.2011.6032878.

[7] 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.

[8] S. Ahmad, and M. Alam "Balanced-Ternary Logic for Improved and Advanced Computing", IJCSIT, 2014, Vol. 5, Issue 4, pp. 5157-5160.

[9] 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.

[10] 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.

[11] D.V. Gavzov, V.V. Sapozhnikov, and Vl.V. Sapozhnikov "Methods for Providing Safety in Discrete Systems", Autom. and Remote Control, 1994, vol. 55, issue 8, pp. 1085-1122.

[12] D.J. Smith, and K.G.L. Simpson "Functional safety: A Straightforward Guide to IEC 61508 and Related Standards", Butterworth-Heinemann; 1st edition (June 26, 2001), 208 p.

[13] E. Fujiwara "Code Design for Dependable Systems: Theory and Practical Applications", John Wiley & Sons, 2006, 720 p.

[14] V.V. Sklyar "Ensuring the Safety of Automated Process Control Systems in Accordance with Modern Standards" (in Russ.), Moscow, Infra-Inzhenerija, 2018, 384 p.

[15] E.S. Sogomonyan, and E.V. Slabakov "Self-Checking Devices and Fault-Tolerant Systems" (in Russ.), Moscow: Radio & Communication, 1989, 208 p.

[16] M. Goessel, V. Ocheretny, E. Sogomonyan, and D. Marienfeld "New Methods of Concurrent Checking", Dordrecht: Springer Science+Business Media B.V., 2008, 184 p.

[17] R. Ubar, J. Raik, 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.

[18] V.V. Sapozhnikov, Vl.V. Sapozhnikov, and D.V. Efanov "Hamming Codes in Concurrent Error Detection Systems of Logic Devices" (in Russ.), St. Petersburg: Nauka, 2018, 151 p.

[19] S.J. Piestrak "Design of Self-Testing Checkers for Unidirectional Error Detecting Codes", Wrocław: Ofiyna Wydawnicza Politechniki Wrocławskiej, 1995, 111 p.

[20] G.C. Cardarilli, S. Pontarelli, M. Re, and A. Salsano "Concurrent Error Detection in Reed-Solomon Encoders and Decoders", IEEE Transıo on Very Large Scale Integration (VLSI) Systems, 2007, Vol. 15, Issue 7. pp. 842-846, doi: 10.1109/TVLSI.2007.899241.

[21] S. Bayat-Sarmadi, and M.A. Hasan "On Concurrent Detection of Errors in Polynomial Basis Multiplication", IEEE Trans. on VLSI Systems, 2007, vol. 15, pp. 413-426, doi: 10.1109/TVLSI.2007.893659.

[22] 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.

[23] D. Efanov, V. Sapozhnikov, and Vl. Sapozhnikov "Generalized Algorithm of Building Summation Codes for the Tasks of Technical Diagnostics of Discrete Systems", Proc. 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.

[24] G. Tshagharyan, G. Harutyunyan, S. Shoukourian, and Y. Zorian "Experimental Study on Hamming and Hsiao Codes in the Context of Embedded Applications", Proc. of 15th IEEE East-West Design & Test Symposium (EWDTS`2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 25-28.

[25] 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 EIConRus, 29 January – 1 February 2018, Moscow, Russia, pp. 1430-1433, doi: 10.1109/EIConRus.2018.8317365.

[26] A.E. Brouwer, H.O. Hamalainen, P.R.J. Ostergard, and N.J.A. loane "Bounds on Mixed Binary/Ternary Codes", IEEE Transactions on Information Theory, 1988, vol. 44, Issue 1 pp. 140-161, doi: 10.1109/18.651001.

[27] 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.

[28] 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.

[29] 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.

[30] 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 Int. Castle Meeting, ICMCTA 2017, Vihula, Estonia, August 28-31, 2017, pp. 228-237.

[31] J.M. Berger "A Note on Error Detecting Codes for Asymmetric Channels", Inf. and Control, 1961, vol. 4, issue 1, pp. 68-73, doi: 10.1016/S0019-9958(61)80037-5.

[32] D.V. Efanov "Ternary Parity Codes: Features", Proc. of 17th IEEE East-West Design & Test Symposium (EWDTS`2019), Batumi, Georgia, September 13-16, 2019, pp. 315-319, doi: 10.1109/EWDTS.2019.8884414.

[33] D.V. Efanov, V.V. Sapozhnikov, and Vl.V. Sapozhnikov "On Summation Code Properties in Functional Control Circuits", Autom. and Remote Control, 2010, Vol. 71, Issue 6, pp. 1117-1123, doi: 10.1134/S0005117910060123.

Quarry Areas Segmentation on Satellite Images by Convolutional Neural Networks

Roman Larionov
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
r.larionov1@uniyar.ac.ru

Vladimir Khryashchev
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
v.khryashchev@uniyar.ac.ru

Vladimir Pavlov
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
vladimir@1pavlov.com

Alexander Ganin
Tochka Zreniya LLC
Yaroslavl, Russia
ganin@tochka.ai

***Abstract*—This article presents results of two deep learning algorithms for sand quarries detection on high-resolution aerial photos. Planet database of high-resolution aerial images was collected. Input images contain blue, green, red and near-infrared channels. Before the training process there was implemented the equalization of brightness histogram which allows to cope with the problem of obscuration of satellite images. To implement numerical experiments there were extracted smaller patches. Training and test sets were enlarged by three types of data augmentation. Convolutional neural networks were pretrained on the SpaceNet dataset and tuned on the Planet database. Deep learning algorithms were launched on NVIDIA DGX-1 supercomputer. Special metrics, such as F1, precision, recall and Dice coefficient allowed to compare the quality of developed models.**

Keywords— quarry areas detection; computer vision; high-resolution aerial photos; convolutional neural networks; Mask R-CNN; U-Net

I. INTRODUCTION

The launch of the Landsat-1 satellite in 1972 ushered in an era of digital remote sensing data. In the last decade, rapid development in space technologies both in satellite sensors and in data processing capabilities has been observed. The growing number of earth observation satellites produce an expanding amount of data, which requires a set of tools to process and extract information from. The effectiveness of remote sensing data processing depends on the quality of developed algorithms [1].

Today, large number of algorithms for detecting objects on satellite images exists. Classical segmentation algorithms are based on either difference or post-classification methods. The first use brightness difference of the corresponding pixels or the distance between them in a multi-dimensional space of features. Post-classification is based on the preliminary classification of multispectral images and the definition of pixels that have changed the class index (interclass transitions). Each approach has a number of limitations and disadvantages. But in the last years we observe a machine learning boom in many applied tasks including processing of remote sensing data.

There are many classical algorithms for satellite images segmentation, in particular the SVM, k-means and histogram methods, etc. [2] But in the second decade of the 21st century, the usage of convolutional neural networks (CNNs) has become a de facto segmentation method. The work of classical methods is based on the use of information about the color of object and the difference in brightness of neighboring pixels. In addition, CNNs allow taking into account information about the spatial position and the surrounding context, and also, unlike the FCN, the orientation of the object does not play a role. So CNNs are capable to detect ships, planes, buildings, deforestation areas etc. in satellite images in real-time.

This paper presents developed CNNs for the task of quarries detection. Tracking the mines boundaries helps to monitor resource usage and assess the environmental situation in the region. Quarries have the following special features:

- Quarries can vary greatly not only in size (from 10 ha and less to 38 ha and more), but also in their degree of flooding and overgrowing: the quarry can be a sandy surface, partially flooded territory, etc.
- Active mining quarries change daily. Accordingly, their shape is random. Examples of sand quarry objects are shown in Fig. 1.
- Sand quarries, as a rule, occupy a large area and are presented in a single amount in the aerial image.
- Quarries are not very different from any other sandy surfaces or even airfields on aerial image. Developed algorithm should use information about the surrounding objects to solve this problem.

This paper presents CNNs that can be used for sand quarries detection on satellite images. The training process, data preparation, testing and special metrics for assessing the quality of neural network work are described. Our work continues research, which was presented in [3].

This article consists of six parts. In the first part, a reader gets acquainted with the formulation of problem. The introduction devotes to CNNs as an approach in machine learning and complications of image segmentation. The second part describes the collected database of satellite images and its preparation for training the neural. The third part refers to proposed deep learning algorithms including description of Mask R-CNN and U-ResNet34 architectures. The fourth part presents the results and analysis of numerical experiments . In the conclusion there is summarized the research. And finally, the last sections contain the acknowledgement and references.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

Fig. 1. Examples of sand quarries

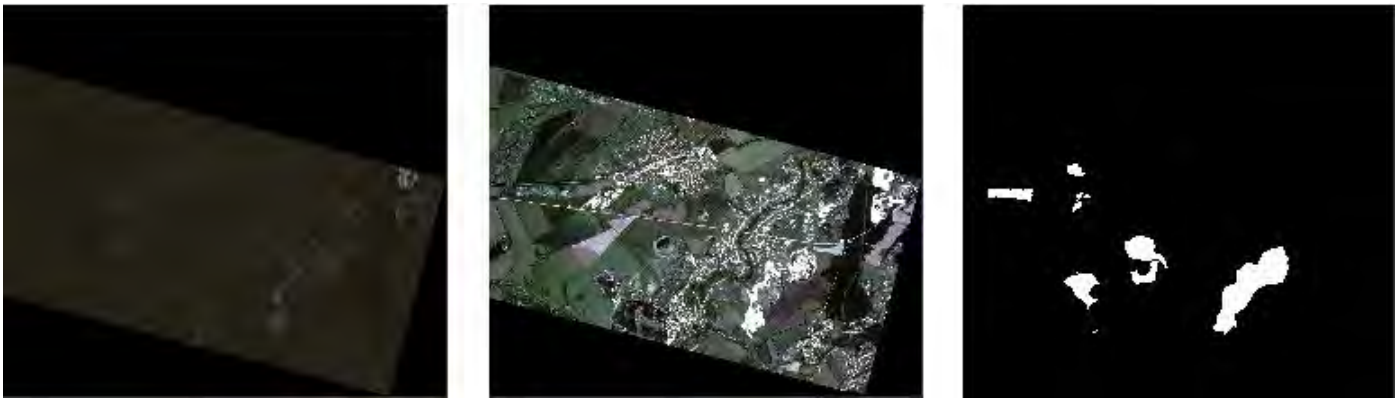
II. Data Preparation

At present time there are several large high-resolution satellite image datasets for training machine learning algorithms. For example, Jilin-1 [4], the SkySat-1 [5], Spacenet [6] and Inria datasets. But these databases have a limited number of manually marked masks. For example, the Inria dataset contains only buildings ground-truth masks. Thus, in our research, we used Spacenet dataset for pre-training developed neural networks, that is, for the better weights initialization.

The main dataset of quarry images was collected by us from RapidEye satellite sensor. 12-bit four-channel (blue, green, red, near-infrared) satellite images from the Planet database have a spatial resolution of 3 m/pixel. Each of 55 high-resolution aerial photos was manually marked by ten different people. Some satellite images from the Planet database have noisy pixels, such as photographed clouds or glares from building roofs and water.

Noisy and very bright pixels may shift the general brightness histogram of image. This side effect leads to the obscuration of full image. It affects the training process of CNNs and as a result a deep learning algorithm might work incorrectly. The equalization of brightness histogram allows to cope with this problem [7].

Histogram equalization is a process of image preprocessing which redistribute all pixels of image uniformly. Values of pixels that do not fall into the certain range of brightness are replaced with minimal or maximal threshold values of brightness. To align the histogram and increase its range of brightness, a linear transformation was performed for each pixel of satellite image. image preprocessing by carrying out histogram equalization significantly reduces the training time of the neural network. Original satellite image, corresponding equalized image and ground-truth mask from the Planet database is shown in Fig. 2.



a) b) c)

Fig. 2. Example of an original image, its equalized copy and expert markup from the Planet database: a) original image, b) equalized image, c) true mask

Since images obtained from different satellite sensors have different bit per pixel, training and testing of neural networks was carried out on normalized images from Planet database. That is, the input images were divided by 4095 and brought into values from the [0, 1]. Satellite images are usually large. Its size can easily exceed 16000 × 16000 pixels. So before the training of CNN by means of data windowing each high-resolution photo and mask of dataset have been sliced on parts of 512 × 512 and 1024 × 1024 size with the step of 256 and 512 respectively. The intersection of patches by half allows to cope with problem of artifacts that appear at the border pixels of image patches due to the peculiarities of the convolution operation. Information about prepared patches of different size is presented in Table I, and Table II.

TABLE I. PREPARED PATCHES OF 512 × 512 SIZE

	Training set	Test set
Total	18900	9060
With objects	1831	652
Without objects	17069	8408

TABLE II. PREPARED PATCHES OF 1024 × 1024 SIZE

	Training set	Test set
Total	12585	1525
With objects	2359	213
Without objects	10226	1312

To increase the training set and increase its diversity, there were used the following stages of augmentation:

- Rotations on π/2, π and 3π/2 and mirroring of corresponding patches. This made it possible to increasy training and test sets by 8 times;
- Applying chromatic distortion;
- Image shifts within 2% of image size, scaling on a coefficient from [1; 1,2] and rotations on small angles from [-15˚, + 15˚].

III. CONVOLUTIONAL NEURAL NETWORKS

In this section there are described developed CNNs used for sand quarries detection on high-resolution aerial photos, and some peculiarities of their training.

First of all, for quarries detection on satellite images there was developed the special neural network architecture called Mask R-CNN [8]. This deep learning algorithm extends the functionality of well-known method for object detection – Faster R-CNN. The architecture of Mask R-CNN is shown in Fig. 3. Quarries are presented in a single amount in the image. Accordingly, it makes no sense to process the entire image, it is better to first detect the areas of interest and then segment them. That's why we turned our attention to Mask R-CNN.

In Mask R-CNN a fully connected neural network (FCN) as an additional branch was added to the basic algorithm. FCN allows to make a segmentation at the pixel level in areas of detected objects (RoI) in parallel with existing outputs of Faster R-CNN: classes and bounding boxes.

Faster R-CNN represents a sequence of two machine learning algorithms: a region-proposal network (RPN) and Fast R-CNN [9]. RPN makes predictions about possible locations of bounding boxes for areas of interest (RoI) on satellite images using sliding windows and «anchors» - special rectangular frames of various size and different ration of sides that surround objects of interest. The presence or absence of objects inside each frame is determined due to the value of IoU metrics («intersection-over-union»): if the value of IoU is more than 0.5 then it is considered that the object fell into the frame. As RPN architecture, we used Feature Pyramid Network (FPN). The main advantage of this approach is to improve the quality of detection, taking into account a wide range of possible sizes of objects. Feature maps of lower and upper layers of FPN have pros and cons: lower layers have high resolution, but low semantic, generalizing ability, whereas upper layers have low resolution, but good generalizing ability to extract needed unobvious features. Fast R-CNN, extracts a set of frames which surround objects of interest more exactly and classifies detected objects simultaneously. The key components of this deep learning algorithm are the spatially pyramidal layer with RoIPool operation, which extracts a feature map for each RoI and RoIAlign layer, which determines the exact spatial location of frames for detected objects on images.

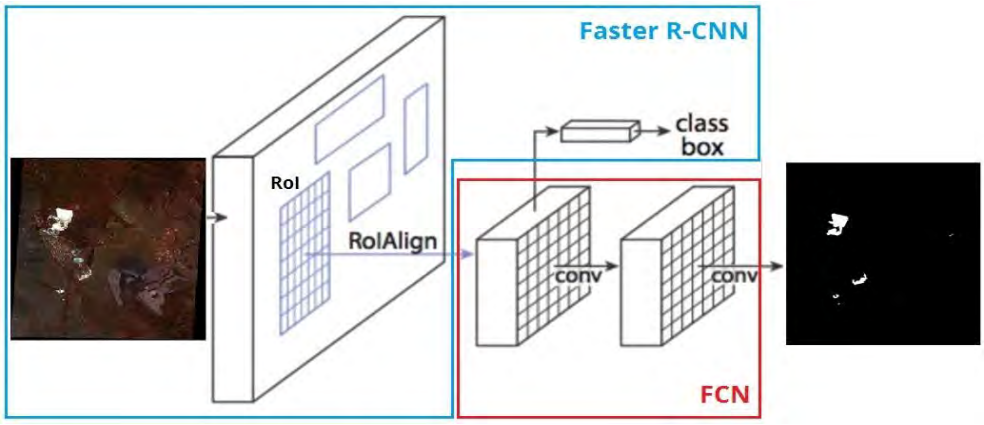


Fig. 3. Architecture of Mask R-CNN

Also there was developed U-Net-like architecture – U-ResNet34 based on models from paper [10]. U-ResNet34 is a U-Net neural network, where ResNet34 used as an encoder and decoder was copied from the classic U-Net architecture.

IV. Numerical Results

The training and testing of Mask R-CNN and U-ResNet34 were carried out on NVIDIA DGX-1 supercomputer. This computing server has 8 GPUs NVIDA Tesla V100 with 16 GB of memory. The peak computing performance is almost 1 petaFlops. Thus, the neural networks training did not exceed 2 days.

All developed algorithms were implemented using Tensorflow framework. We used Adam with learning rate of 1e-3 to optimize the training process. This algorithm halp to avoid early fallings of the loss function to local minima during gradient descent training. The value 1e-3 was chosen experimentally. At lower values, the loss function changes too slowly and the probability of overfitting significantly increases.

To assess the detecting and segmenting ability of an algorithm, special metrics from digital image processing are often used. In our research there was used IoU coefficient, a binary measure of similarity predicted and true masks. Also to measure the quality of developed deep learning algorithms there were used precision (P), recall (R) and F-score (F_1). During the training, on the assumption of maximal value of F_1, for each CNN there was chosen a threshold value to form predicted masks with detected sand quarries. If the value of possibility for a pixel is higher than the threshold value, it belongs to the class of interest.

The correct choice of the loss function allows taking into account all the features objects of interest. In our research there was used a sum of binary cross entropy (BCE) and the value of Dice loss (DL) based on Sorensen coefficient [11]:

$$Loss = BCE(X,Y) + DL(X,Y),$$

$$BCE(X,Y) = -\sum_{x,y}(xlog(y) + (1-x)log(1-y)),$$

$$DL\,(X,Y) = 1 - \frac{2|X \cap Y|}{|X| + |Y|}.$$

Quarries vary in size greatly, from tens to hundreds of meters in linear dimensions. Therefore, it is necessary to choose the right sizes for anchors to cover small and large objects of interest. For Mask R-CNN there were selected two sets of anchors:

- Anchor set 1 (AS 1): 64, 128, 256, 512 and 1024;
- Anchor set 2 (AS 2): 32, 64, 128, 256 and 512.

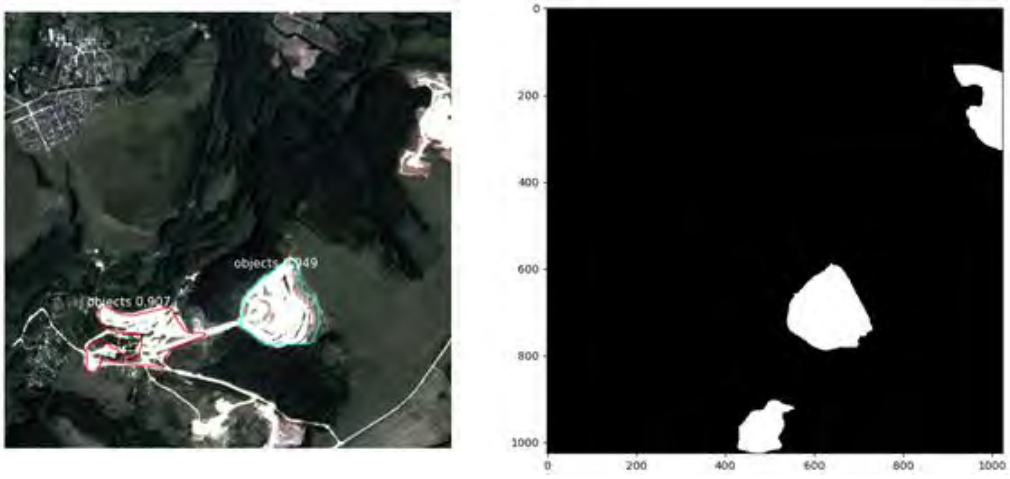
For Mask R-CNN there were used training and test sets of patches of 1024 × 1024 size which were enlarged using all techniques described earlier. The training and tuning processes finish after completing 113 epochs. Test results for Mask R-CNN with different sets of anchors on the Planet database were presented in Table III.

Mask R-CNN with the set of larger anchors is shown better results: the value of F_1 reached 0,252 in comparison with 0,142 which was given on the set of smaller anchors. This peculiarity can be explained by the fact that sand quarries are big enough on satellite images so the deep learning algorithm with little bounding boxes are not able to detect whole objects of interest. The example of an input image with detected objects and corresponding true mask of the Planet database for Mask R-CNN is shown in Fig. 4.

The second model, U-ResNet34, was trained on two sets of patches of 512 × 512 size. Sets of fragments were enlarged by methods of data augmentation in two ways:

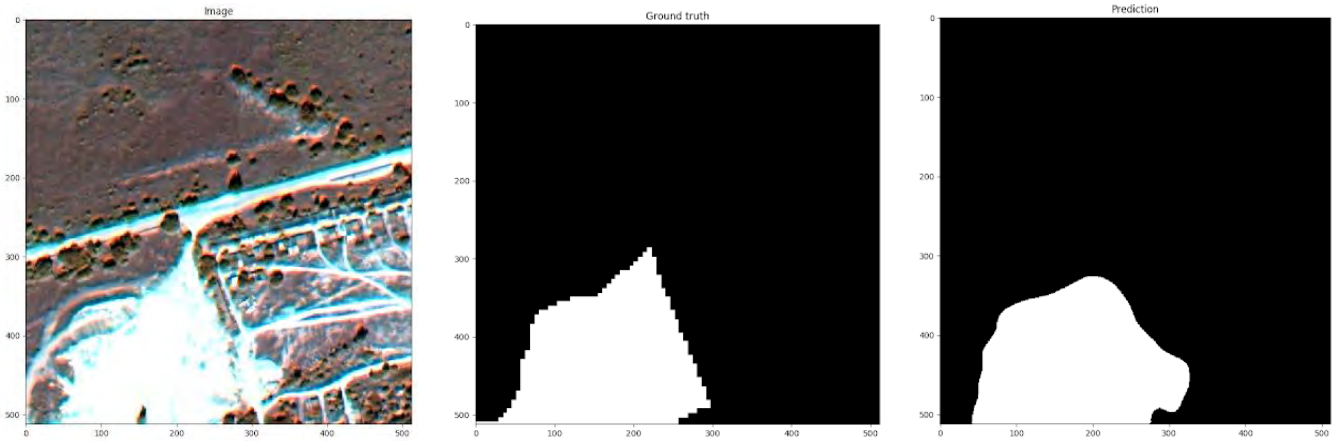
1. Only flips.
2. Flips and SSR.

U-ResNet34 training process with a batch of 16 samples ended after 80 epochs. The values of metrics after testing are presented in Table IV.



a) b)

Fig. 4. Test results for Mask R-CNN: a) input image with detected objects, b) true mask



a) b) c)

Fig. 5. Test results for U-ResNet34: a) input image, b) true mask, c) predicted mask

TABLE III. TEST RESULTS FOR MASK R-CNN ON THE PLANET DATABASE.

Metrics	Mask R-CNN (AS 1)	Mask R-CNN (AS 2)
IoU	0,753	0,757
F_1	0,252	0,142
P	0,189	0,081
R	0,380	0,584

TABLE IV. TEST RESULTS FOR DEVELOPED MODELS ON THE PLANET DATABASE.

Metrics	U-ResNet34 (flips)	U-ResNet34 (flips + SSR)
IoU	0,772	0,765
F_1	0,408	0,357
P	0,375	0,289
R	0,447	0,465

According to results presented in Table III and Table IV, the best value of F_1 for U-ResNet34 reached 0,408 and exceed the same metric for Mask R-CNN on 0,156. Thus, U-ResNet34 works better in the task of quarries detection on high-resolution aerial photos. However, the usage of image shifts and rotations on small angles do not improve the quality of deep learning algorithm: an amount of false positive objects increased vastly. Small sandy areas usually acted as false positive objects. The example of an input image, true and predicted masks of the Planet database for U-ResNet34 is shown in Fig. 5.

An experiment on processing time of one satellite image was also carried out. The input 4-channel image had shape of 16384 × 16384. U-ResNet34 showed the best results so we chose it for the test. The input image was divided into patches size of 512 × 512 pixels, and then the patches were glued together. Processing was carried out on one GPU NVIDA Tesla V100. The average processing time for one patch was about 19 ms and the total image was processed in less than 20 seconds.

V. CONCLUSION

This article presents numerical experiments for developed deep learning algorithms: Mask R-CNN and U-ResNet34. For training and testing, the Planet database of satellite images was collected and marked manually. Before the training process there was implemented the equalization of brightness histogram which allows to cope with the problem of obscuration of satellite images. To implement numerical experiments there were extracted smaller patches. Training and test sets were enlarged methods using various methods of data augmentation. The developed algorithms were pretrained on the SpaceNet dataset and tuned on the Planet database. According to test results U-ResNet34 works better in the task of sand quarries detection on high-resolution aerial photos. Its average values of IoU and F_1 are equal 0.772 and 0.408 respectively. In further research, we will try to process satellite stereo images and calculate the volume of detected quarries. It gives more control over the extraction of natural resources.

VI. REFERENCES

[1] J. Liu and P. Mason "Image processing and GIS for remote sensing: techniques and applications", John Wiley & Sons, 2016, p. 472.

[2] D. Kasimov, A. Kuchuganov and V. Kuchuganov, "Methods and Tools for Developing Decision Rules for Classifying Objects in Aerial Images," in Proc. FRUCT'26 conference, 2020, pp. 158-165, doi: 10.23919/FRUCT48808.2020.9087419.

[3] R. Larionov and V. Khryashchev, "Wildfire Segmentation on Satellite Images using Deep Learning," in Proc. *Moscow Workshop on Electronic and Networking Technologies (MWENT)*, 2020, pp. 1-5. doi: 10.1109/MWENT47943.2020.9067475

[4] X. Tong, G. Xia, Q. Lu and H. Shen, "Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models," Web: https://arxiv.org/pdf/1807.05713.pdf

[5] SkySat-1 Satellite Images. Web: https://www.satimagingcorp.com/satellite-sensors/skysat-1/.

[6] SpaceNet Database. Web: http://explore.digitalglobe.com/spacenet.

[7] R. Dorothy, R. Joany and J. Rathish, "Image enhancement by Histogram equalization," International Journal of Nano Corrosion Science and Engineering, vol. 2, 2015, pp. 21-30.

[8] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," in Proc. IEEE International Conference on Computer Vision, 2017, pp. 2980-2988. doi: 10.1109/ICCV.2017.322

[9] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," I*EEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149. doi: 10.1109/TPAMI.2016.2577031

[10] A. Buslaev, S. Seferbekov, V. Iglovikov and A. Shvets, "Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery," Web: https://arxiv.org/pdf/1806.05182.pdf

[11] R. Larionov, V. Khryashchev and V. Pavlov, "Separation of Closely Located Buildings on Aerial Images Using U-Net Neural Network," in Proc. FRUCT'26 conference, 2020, pp. 256-261. doi: 10.23919/FRUCT48808.2020.9087365

Antenna Arrays Calibration Using Recursive Least Squares Adaptive Filtering Algorithms Based on Inverse QR Decomposition

Victor Djigan
Department of Integrated Circuits Design Methodology
Institute for Design Problems in Microelectronics of RAS
Moscow, Russia
djigan@ippm.ru

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

***Abstract*—This paper considers the antenna arrays calibration by the using of the Recursive Least Squares (RLS) adaptive filtering algorithms. An algorithm, based on the inverse QR decomposition, has been selected among the diversity of the RLS algorithms. This is caused by its stable operation. Because the algorithm contains the computationally heavy square root operations, a square root free version of the algorithm is also presented. Both versions of the QR RLS algorithms are mathematically identical to each other if they operate in float point arithmetic. The proposed calibration can be used in the antenna arrays with digital beamforming, because the algorithms usage requires the access to the array channel signals. The calibration requires a known training signal, which can be easily provided not only in a laboratory environment, but also in a field operation, if an array is used as a directional antenna of the digital communication system equipment. In the second case, the calibration can be also conducted even in the presence of the interference signal sources. Simulation validates the proposed calibration algorithm, using linear antenna arrays with 4, 8 and 16 antennas with a half wavelength distance between the neighbor antennas. In this simulation, the array channel noise has been varied in 0 … 30 dB range of the Signal-to-Noise Ratio. Two interference sources with the –30 dB Signal-to-Interference Ratio each have been simulated. These sources were located symmetrically relatively the required main lobe direction of the array radiation pattern. A training signal has been simulated as a random one with no specific autocorrelation properties. The signal has been modulated by the Phase Shift Keying (PSK) and the Quadrature Amplitude Modulation.**

Keywords—Antenna array, digital beamforming, array calibration, Recursive Least Squares (RLS), QR decomposition

I. INTRODUCTION

An Antenna Array (AA) is a sort of directional antennas, which is often used in the equipment of modern telecommunication systems. Today there is a diversity of the AA configurations: linear, flat, circular, cylindrical or conformal. This configuration depends on the AA application, the installation conditions and the required shape of the Radiation Pattern (RP) [1].

Usually, each channel of an AA contains a variable weight, that may be a digitally controlled phase shifter and/or an attenuator. These weights ensure the required shape of RP over the range of the observing angles. The RP shaping is provided by the calculation (synthesis) of these weights. The calculation is based on using the ideal AA model.

However, due to the variation of the materials properties, active and passive radio frequency (RF) component parameters, after the manufacturing the AA might have the characteristics, which are far from the expected ones. And this does not allow to use the RP shaping correctly.

To overcome this problem, the AA are usually calibrated after the manufacturing and/or during the field operations. The calibration assumes the estimation of the deviation of the complex-valued channel gains from the ideal ones, specified by the AA design, and taking these estimates into the consideration (means compensation) during the AA field operation.

The diversity of the AA calibration methods and algorithms can be found in [2]. Some new ones are also presented in [3], [4]. Most of the calibration algorithms are developed for the using in the traditional scanning AA, which do not provide the access to the AA channel signals.

However, due to the interest to the adaptive AA technology [5], which is based on the modern adaptive signal processing [6], [7], a new sort of the AA became commercially available. These are the AA with Digital Beam Forming (DBF) [8]. The DBF AA manufacturing became possible due to the achievements in modern microelectronics, that allowed to manufacture a wide range of digital integrated circuits, used together with the active and the passive RF components in design of the AA blocks and their signal processing units.

The DBF AA provides the Base Band (BB) channel signal access that is one of the requirements of signal processing algorithms in the most of the adaptive AA. This access also provides an additional degree of freedom for the DBF AA calibration. This means the following: in the case of the DBF AA the calibration can be conducted using the same adaptive filtering algorithms as the ones, used for the interference signal

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

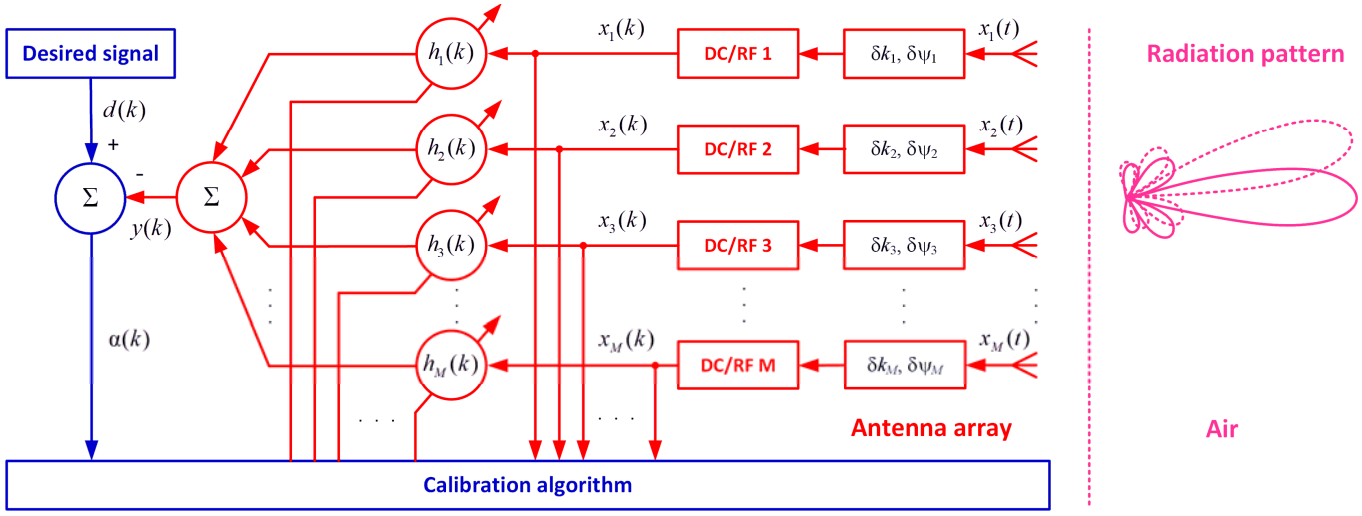


Fig. 1. AA calibration based on adaptive signal processing: a generalized architecture

cancellation. However, these algorithms are not used to minimize the AA response towards the interference sources, but to maximize the AA response towards the desired signal source locations.

This paper considers the usage of the Recursive Least Squares (RLS) adaptive filtering algorithms, based on the inverse QR decomposition, for the calibration of the AA with DBF. The paper presents two sorts of the algorithms: with and without square root operations.

The simulation examples validate the proposed solution of the calibration problems for the AA with a different number of antennas in the interference and interference free conditions and in a wide range of the AA channel Signal-to-Noise Ratio (SNR).

II. CALIBRATIUON PROBLEM AND ITS SOLUTION

The AA with DBF contains the RF modules, which receive, select and amplify all signals accepted by the antennas, and contains the Down Converters (DC), which transfer the AA channel RF signals to the Base-Band (BB) ones, see Fig. 1. The BB signals are sampled at Nyquist rate, weighed by the multiplication with complex-valued weights and combined in digital domain, producing DBF AA.

The AA, see Fig. 1, has M antennas/channels. The complex-valued gain of a part of the AA channel form from the m-th antenna input to the m-th weight h_m is not the same in each channel. This is caused by the channel phase $\delta\psi_m$ and real-valued gain δk_m errors, where $m = 1, 2, 3, ..., M$.

Continuous time t signals, received by the AA antennas, compose the signal vector

$$\mathbf{x}_M(t) = \left[x_1(t), x_2(t), \ldots, x_m(t), \ldots, x_M(t)\right]^{\mathrm{T}}, \qquad (1)$$

which after the DC is transformed to the discrete time signal vector

$$\mathbf{x}_M(k) = \left[x_1(k), x_2(k), \ldots, x_m(k), \ldots, x_M(k)\right]^{\mathrm{T}}, \qquad (2)$$

where k is the number of the processed signal samples. Usually, these samples are coincided with the signal processing iterations.

Through the paper, the lowercase and uppercase characters denote the scalar variables and the elements of the vectors and matrices. The vectors and the matrices are denoted by the bold lowercase and uppercase characters, respectively. The superscript T denotes the transposition of a vector or a matrix, the superscript H denotes the Hermitian transpose, i.e. transposition of a vector or a matrix and the complex conjugation of its elements, denoted as $*$. The subscript M indicates the number of the elements in a vector or the number of the elements $M \times M$ in a square matrix. The subscripts of the scalar variables denote the numbers of an element in a vector or in a matrix.

The objective of the AA calibration is to find a vector of its weights, see Fig. 1,

$$\mathbf{h}_M(k) = \left[h_1(k), h_2(k), \ldots, h_m(k), \ldots, h_M(k)\right]^{\mathrm{T}}, \qquad (3)$$

which has to satisfy the following condition:

$$\mathbf{h}_M(k) \otimes \mathbf{k}_M = \mathbf{c}_M^{*}. \qquad (4)$$

In (4)

$$\mathbf{k}_M = \left[(1+\delta k_1)e^{j\delta\psi_1}, (1+\delta k_2)e^{j\delta\psi_2}, \ldots, (1+\delta k_m)e^{j\delta\psi_m}, \ldots, (1+\delta k_M)e^{j\delta\psi_M}\right]^{\mathrm{T}} \qquad (5)$$

is the vector of the complex-valued AA channel gains, which values are unknown. The symbol $\otimes$ denotes the operation of multiplication of the elements with the same numbers in two vectors.

The vector

$$\mathbf{c}_M = \left[e^{j\psi_1(\theta_s)}, e^{j\psi_2(\theta_s)}, \ldots, e^{j\psi_m(\theta_s)}, \ldots, e^{j\psi_M(\theta_s)}\right]^{\mathrm{T}} \qquad (6)$$

is the so-called steering vector. Here, the $\psi_m(\theta_s)$ are space phase lags, which depend on the geometrical configuration of the AA, a reference antenna location and m value [1].

In an ideal AA, the weights vector, selected as

$$\mathbf{h}_M(k) = \mathbf{c}_M^*, \quad (7)$$

ensures the AA main lobe of RP steering towards the θ_s direction. This direction is counted relatively the broadside direction (the normal to the AA aperture).

To calculate the vector $\mathbf{h}_M(k)$, which satisfies the (4) condition, it is suggested to use the adaptive filtering algorithms, which minimize some function of the errors

$$\alpha(k) = d(k) - y(k) \quad (8)$$

between the training $d(k)$ and the AA output $y(k)$ signals.

The function minimization ensures the maximization of the AA response towards the θ_s direction. The task has a unique solution in the case of the quadratic cost function, independently of the $\mathbf{k}_M$ and $\mathbf{c}_M$ vector values.

If AA receives only one training signal, the AA input signal correlation matrix

$$\mathbf{R}_M(k) = E\left\{\mathbf{x}_M^{\mathrm{H}}(k)\mathbf{x}_M(k)\right\}, \quad (9)$$

is a singular [7], where $E\{\bullet\}$ denotes the expectation.

Due to this reason, the solution of the AA calibration task by means of the simple linear $O(M)$ arithmetic complexity gradient search based adaptive filtering algorithms is not efficient. The algorithms behavior depends on the correlation matrix eigenvalues spread.

Other adaptive filtering algorithms, which do not suffer of the above drawback, are the RLS ones, which in case of the usage in architecture, see Fig. 1, have the quadratic arithmetic complexity $O(M^2)$. The algorithms behavior does not depend on the correlation matrix eigenvalues spread [6].

There is a number of such adaptive filtering algorithms, among which the most frequently used are the algorithms, based on Matrix Inversion Lemma (MIL), inverse QR decomposition and Householder transform [6].

The RLS algorithms, based on QR decomposition, are characterized by a stable operation, which is an important feature for the processing of the correlated signals with the high spread of the correlation matrix eigenvalues. Due to the reason, these algorithms have been selected for the research, presented in this paper.

The basic version of the algorithm contains the square root operations, which require additional computational recourses for the function implementation. There is also a mathematically identical (in case of float point arithmetic) version of the algorithm, which has no square root operations. These both RLS algorithms, fitted to the architecture, see Fig. 1, are presented below.

Calibration algorithm, based on inverse QR decomposition RLS adaptive filtering algorithm

Initialization : $\mathbf{x}_M(0) = \mathbf{0}_M$, $\tilde{\mathbf{R}}_M^{-\mathrm{H}}(0) = \delta^{-1}\mathbf{I}_M$, $\mathbf{h}_M(0) = \mathbf{0}_M$

For $k = 1, 2, \ldots, K$

$\mathbf{x}_M(k) = \left[x_1(k), x_2(k), \ldots, x_m(k), \ldots, x_M(k)\right]^{\mathrm{T}}$

$y(k) = \mathbf{h}_M^{\mathrm{H}}(k-1)\mathbf{x}_M(k)$

$\alpha(k) = d(k) - y(k)$

$u_{M,j}^{(j-1)*}(k) = 0, \quad 1 \le j \le M, \quad b_M^{(0)}(k) = 1$

For $i = 1, 2, \ldots, M$

$$a_{M,i}(k) = \lambda^{-0.5}\tilde{\mathbf{R}}_M^{-\mathrm{H}}(k-1)\Big|_{i,1:i}\ \mathbf{x}_M(k)\Big|_{1:i}$$

$$b_M^{(i)}(k) = \sqrt{\left[b_M^{(i-1)}(k)\right]^2 + a_{M,i}^*(k)a_{M,i}(k)}$$

$$s_{M,i}(k) = \left[b_M^{(i)}(k)\right]^{-1} a_{M,i}^*(k)$$

$$c_{M,i}(k) = b_M^{(i-1)}(k)\left[b_M^{(i)}(k)\right]^{-1}$$

For $j = 1, 2, \ldots, i$

$$\tilde{R}_{M,ij}^{-\mathrm{H}}(k) = c_{M,i}(k)\lambda^{-0.5}\tilde{R}_{M,ij}^{-\mathrm{H}}(k-1) - s_{M,i}^*(k)u_{M,j}^{(i-1)*}(k)$$

$$u_{M,j}^{(i)*}(k) = s_{N,i}(k)\lambda^{-0.5}\tilde{R}_{M,ij}^{-\mathrm{H}}(k-1) + c_{M,i}(k)u_{M,j}^{(i-1)*}(k)$$

End for j

End for i

$$\mathbf{g}_M(k) = \mathbf{u}_M^{(M)}(k)\left[b_M^{(M)}(k)\right]^{-1}$$

$$\mathbf{h}_M(k) = \mathbf{h}_M(k-1) + \mathbf{g}_M(k)\alpha_M^*(k)$$

End for k

Calibration algorithm, based on square root free inverse QR decomposition RLS adaptive filtering algorithm

Initialization : $\mathbf{x}_M(0) = \mathbf{0}_M$, $\bar{\mathbf{R}}_M^{-\mathrm{H}}(0) = \delta^{-1}\mathbf{I}_M$, $\mathbf{h}_M(0) = \mathbf{0}_M$

$\mathbf{K}_M^R(0) = \mathbf{I}_M$

For $k = 1, 2, \ldots, K$

$\mathbf{x}_M(k) = \left[x_1(k), x_2(k), \ldots, x_m(k), \ldots, x_M(k)\right]^{\mathrm{T}}$

$y(k) = \mathbf{h}_M^{\mathrm{H}}(k-1)\mathbf{x}_M(k)$

$\alpha(k) = d(k) - y(k)$

$\overline{u}_{M,j}^{(j-1)*}(k) = 0, \quad 1 \le j \le M, \quad K_B^{B(0)}(k) = 1$

For $i = 1, 2, \ldots, M$

$$\overline{a}_{M,i}(k) = \overline{\mathbf{R}}_M^{-\mathrm{H}}(k-1)\left|_{i,1:i} \mathbf{x}_M(k)\right|_{1:i}$$

$$K_M^{B(i)}(k) = K_M^{B(i-1)}(k) + \lambda^{-1} K_{M,i}^R(k-1)\overline{a}_{M,i}^*(k)\overline{a}_{M,i}(k)$$

$$\overline{s}_{M,i}(k) = \lambda^{-1} K_{M,i}^R(k-1)\overline{a}_{M,i}^*(k) / K_M^{B(i)}(k)$$

$$\overline{c}_{M,i}(k) = K_M^{B(i-1)}(k) / K_M^{B(i)}(k)$$

For $j = 1, 2, \ldots, i$

$$\overline{R}_{M,i,j}^{-\mathrm{H}}(k) = \overline{R}_{M,ij}^{-\mathrm{H}}(k-1) - \overline{a}_{M,i}(k)\overline{u}_{M,j}^{(i-1)*}(k)$$

$$\overline{u}_{M,j}^{(i)*}(k) = \overline{s}_{M,i}(k)\overline{R}_{M,ij}^{-\mathrm{H}}(k-1) + \overline{c}_{M,i}(k)\overline{u}_{M,j}^{(i-1)*}(k)$$

End for j

$$K_{M,i}^R(k) = \lambda^{-1} K_{M,i}^R(k-1)\overline{c}_{M,i}(k)$$

End for i

$$\mathbf{g}_M(k) = \overline{\mathbf{u}}_M^{(M)}(k)$$

$$\mathbf{h}_M(k) = \mathbf{h}_M(k-1) + \mathbf{g}_M(k)\alpha_M^*(k)$$

End for k

The parameter δ^2 of the initial regularization of the correlation matrix $\mathbf{R}_M(k)$ is selected as

$$\delta^2 \ge 0.01\sigma_x^2 \qquad (10)$$

where σ_x^2 is the AA input signal variance. Other details of the above algorithms can be found in [6]. An additional improvement of any RLS algorithm stable behavior can be achieved by using the dynamic regularization. The regularization example concerning MIL RLS algorithm is shown in [6]. The regularization increases the arithmetic complexity of the RLS algorithms about two times. This is the cost, paid for the algorithm stability.

III. Algorithm verification

The considered AA algorithm has been verified via the simulation of the calibration of the linear AA with the omnidirectional antennas, with a half wavelength distance between the neighbor antennas and number of antennas of $M = 4$, $M = 8$ and $M = 16$. The AA channel phase errors have been simulated as the random ones uniformly distributed in the range of $\delta\psi_m = -\pi \ldots \pi$. The AA real-valued channel gain errors δk_m have also been uniformly distributed over the range of 0 … –3 dB. These test-cases demonstrate the AA error estimation and compensation with the accuracy of about –20 dB in terms of norm of the estimated channel gain vector. About – 20 dB accuracy is also demonstrated in terms of the norm of the calibrated and the required RP difference. The results are valid for the array channel SNR=0 … 30 dB and the using of any Phase Shift Keying (PSK) or Quadrature Amplitude Modulated (QAM) training signals with 2 …. 16 element data alphabets.

The examples of the AA calibration in the terms of the RP shapes are shown in Fig. 2. The left pictures in this figure show the results of the AA calibration in the absence of the interference sources and the right ones show the calibration results in the presence of two interference sources, located symmetrically relatively the calibrated AA RP main lobe direction. The Signal-to-Interference Ratio for each interference has been selected as –30 dB. The blue curves of Fig. 2 show the initial RP of an ideal AA. The magenta curves show that the initial RP are completely destroyed by the channel gain errors. The green curves show the required RP and the red ones show the RP of the calibrated AA. The vertical red down arrows indicate the required RP main lobe directions. The vertical blue down arrows indicate the ideal AA RP main lobe initial directions and the directions towards the interference sources, if the sources are present. The figure demonstrates that the calibrated AA RP are close to the required ones. This mean, that during the operation, the calibration algorithm estimates the weight vector $\mathbf{h}_M(k)$, which satisfies the condition (4). The estimation does not depend on the values of the errors in channel gains, the initial θ_i and required θ_s RP main lobe directions. It also does not depend on the existence or absence of the interference sources, because during the calibration the interference sources signals are suppressed by means of the calibration algorithm at the AA output.

IV. Conclusion

Thus, in the paper we have demonstrated the ability of the adaptive filtering RLS algorithms to calibrate the AA in the wide range of the channel SNR and in the presence or absence of the interference sources. This means that the calibration does not require special conditions like an anechoic chamber. Therefore, it can also be conducted in usual laboratories or even during an AA field operation.

References

[1] R.J. Maillou, Phased array antenna handbook, 3rd ed. Artech House, Inc., 2017, 506 p.

[2] E.V. Korotetsky, A.M. Shitikov, and V.V. Denisenko, “Phased antenna array calibration methodths,” Radioengineering, pp. 95–104, May 2013. (in Russian)

[3] V.V. Kurganov, “Antenna array complex channel gain estimation using phase modulators,” Antennas Design and Measurement International Conference. Saint Petersburg, pp. 126–129, 2019.

[4] V.I. Djigan and V.V. Kurganov, “Antenna array calibration algorithm without acces to channel signals,” Radioelectronics and Communications Systems, vol. 61, p. 1–14, January 2020.

[5] B. Allen and M. Ghavami, Adaptive array systems. Fundamentals and applications. John Wiley & Sons Ltd., 2005, 250 p.

[6] V.I. Djigan, Adaptive filtering: theory and algorithms. Moscow: Technosphera Publisher, 2013, 528 p. (in Russian)

[7] S. Haykin, Adaptive filter theory, 5th ed. Pearson Education Inc., 2014, 889 p.

[8] C. Fulton, M. Yeary, D. Thompson, J. Lake, and A. Mitchell, “Digital phased arrays: challenges and opportunities,” Proceedings of the IEEE, vol. 104, p. 487–503, March 2016.

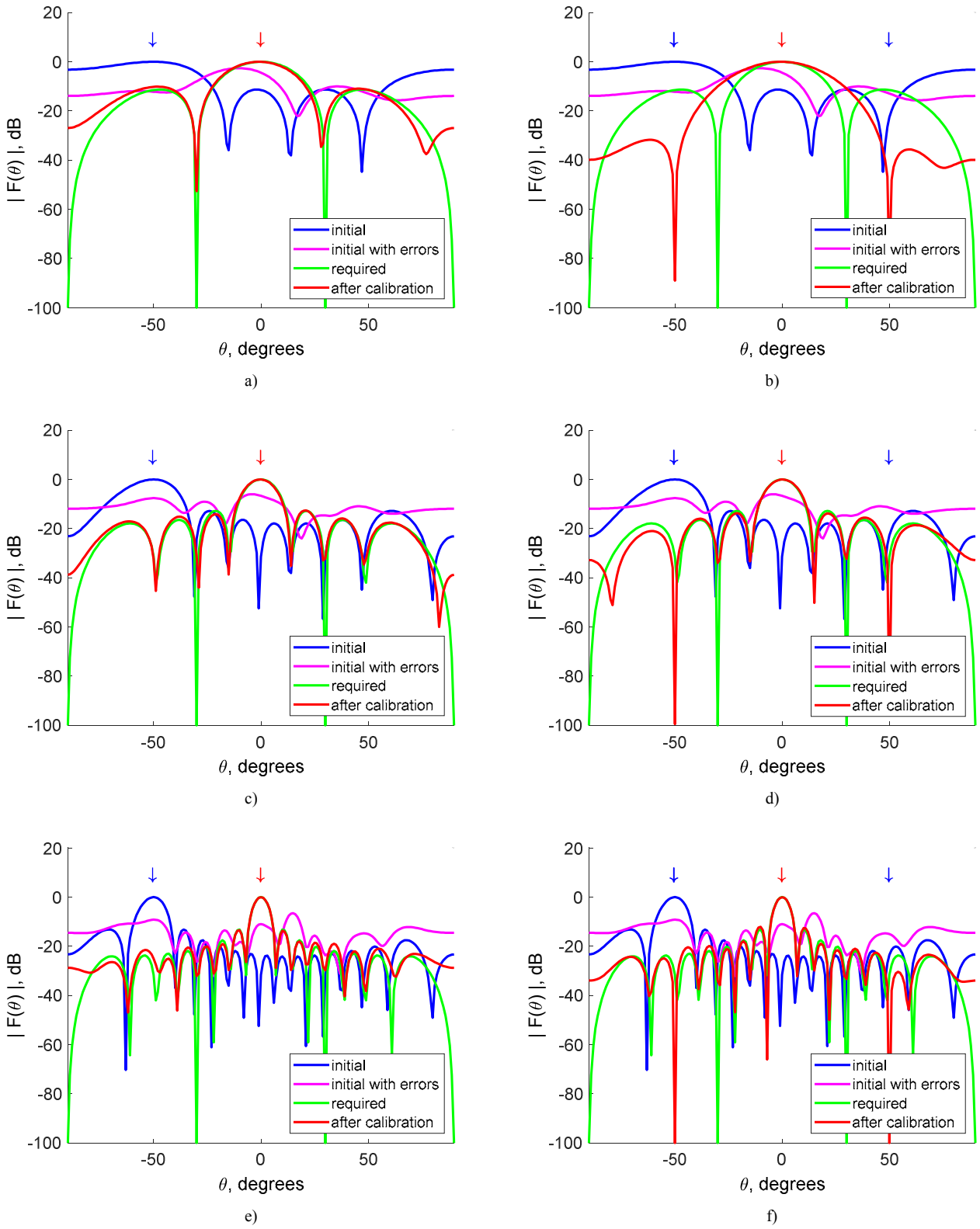


Fig. 2. RP before and afrer AA calibration: a), c) and e) are the cases with no interfernces; b) , d) and f) are are the cases with interfernces

Efficient FPGA Implementation of Field Oriented Control for 3-Phase Machine Drives

Burak Tufekci[1,3], Bugra Onal[2,3], Hamza Dere[2,3], and H. Fatih Ugurdag[2]

[1]CS Dept., Ozyegin University
[2]EE Dept., Ozyegin University
[3]Semimobility Teknoloji Ltd.
Istanbul, Turkey
burak.tufekci.17309@ozu.edu.tr

***Abstract*—This paper presents an FPGA implementation of Field Oriented Control (FOC) method with high switching frequency for 3-phase machine drives. A common architecture has been constructed for both BrushLess DC motors (BLDC) and Permanent Magnet Synchronous Motors (PMSM). For this purpose, the controller module has been implemented by using a hardware efficient algorithm, namely, Coordinate Rotation Digital Computer (CORDIC). The result of this implementation has been compared with the literature, and we claim that this paper's FPGA design has better performance in terms of area and speed with respect to other FPGA-based FOC designs.**

***Index Terms*—Motor Control, FPGA, FOC, BLDC, PMSM, CORDIC**

I. Introduction

Developments in semiconductor technology lead to energy-efficient and high frequencies power switches [1]. Such devices enabled us to come up with high frequency power system design solutions. High-frequency approach has many benefits in 3-phase (3-ϕ) motor drive applications [2]. Among these benefits can be better motor efficiency, low-cost filtering, lower torque ripple and faster control response. While higher frequencies of Pulse Width Modulation (PWM) have these advantages, they also cause voltage reflection and motor insulation breakdown issues at the motor terminals. Therefore, the operating PWM frequency and the type of the motor must be examined carefully.

Increasing the PWM frequency may not be easy for any setup. While the operating frequencies of microcontrollers are sufficient to exceed our device frequencies, it does not change the fact that microcontrollers can sometimes reboot and sequential iteration process can often take too long to measure non-linear operations. On the other hand, algorithm development on microcontrollers is faster than other semiconductor platforms.

FPGAs are superior to microcontrollers in many areas in terms of latency, connectivity, and energy consumption. Latencies of FPGA implementations can be 1 millisecond or even less, while even with the best CPUs introduce latency of approximately 50 milliseconds. Furthermore, because FPGAs do not contain any cache or OS, the delays are deterministic. Because input and output can be directly connected to FPGAs, this can enable high bandwidth implementations. FPGAs pin voltages are usually adjustable, they are very good at minimizing energy consumption.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

Since both the speed control range of BrushLess DC motor (BLDC) and Permanent Magnet Synchronous Motor (PMSM) is large and the energy losses are lower in high-frequency applications compared to other motor types, PMSM and BLDC provide an ideal environment for the testing purposes of this work. Besides, FPGAs in a high speed design can respond better and FPGAs control mechanism is safer, they may be a good alternative for testing.

In the literature, Kung et al. proposed an FPGA-based approach to speed control with FOC [3]. An Sliding Mode Observer (SMO) design has been implemented in their work using a sensorless FOC and a phase-locked loop. The speed information has been generated by the user using the NIOS II processor and all other topologies have been implemented in the FPGA. According to the results obtained by the authors, the back-emf graphics in the transition from stop to acceleration can be smoothed by their own approach. Suneeta et al. have introduced FPGA-based control of 3-ϕ BLDC [4]. It has been shown to be more powerful and safer than microcontroller-based electric motor control because of the high design freedoms offered by FPGA-based electric motor controls. Otherwise, because of faster design development, microcontroller-based control is more powerful than FPGA-based controller, and is also cheaper than FPGA-based controller. Hence the choice of controller based on FPGA or microcontroller based control depends on system requirements.

Babu and Athul have used the PI-controller viewpoint to execute FOC on asynchronous motor [5]. They have built their architecture on Xilinx Virtex-5 using the Xilinx System Generator (XSG) toolbox. Since they've used XSG toolbox when implementing this control system on FPGA, we can't say exactly that their architecture is efficient in terms of memory space and maximum clock speed. On the other hand, it will not change the fact that they have done great research by comparing Direct FOC and Indirect FOC approaches. Joakim Eriksson et al. have researched a rapid prototyping system for 3-ϕ electric motor systems [6]. They concluded that a multi-axis device can be rendered with FOC using FPGA. In their work, they have verified nominal torque values, nominal

power values and rated speed values. Besides, the PWM frequency tests have been analyzed. They have found in the simulation findings that the lower current fluctuates at large PWM frequencies. They concluded that as the PWM frequency rises slowly, the electrical motor currents are beginning to deform due to the fact that the Metal-Oxide Semiconductor Field Effect Transistor (MOSFET) may not have enough time to turn on and off entirely during the switching pulses. These problems have been quite overcome in Silicon Carbide (SiC) and Gallium Nitride (GaN) based power switches [7].

Marufuzzaman et al. have suggested a new dq PI controller focused on FPGA [8]. They argued that the new dq PI controller is the main element in increasing the overall output of the system. No matter how correct they are in their paper, there are a lot of performance criteria beyond that. Akin et al. have researched indirect control of the FPGA Induction Machine (IM) [9]. The Vector Control method has been investigated and claimed that although the DTC method is used regardless of the motor parameters, efficient feedback with DTC at low speeds can not be achieved. They therefore thought that the FOC approach would be more effective than the DTC and they have prepared FOC using the XSG toolbox on Xilinx Spartan-3.

In section II, the theory behind FOC has been mentioned. The hardware implementation and the result of the hardware implementation has been analyzed and implementation result has been discussed in section III.

II. Field Oriented Control Methodology

FOC is a Variable Frequency Drive (VFD) control methodology. In Fig. 1, based on W_{REF} value which is rotational speed command, θ value which is coming from Encoder (could be Resolver, Hall, or Sensorless mechanism), and W_{ACT} value that is calculation of speed with respect to given θ value, the circuit tries to reach W_{REF} value by applying necessary steps.

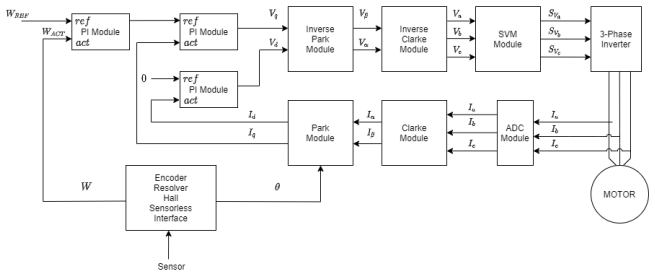


Fig. 1. FOC Flow Graph

The measurement of the motor's rotational speed is a challenge because the running motor has many disturbance factors. The main idea behind the FOC is to make more realistic observations by changing the observation frame to measure motor speed. As a result, our observation has been getting closer to real values, and driving the motor has become more stable.

In any 3-ϕ motor, the sum of 3-ϕ voltages or currents at any time should be equal to zero. By using this approach, transforming voltages or currents between the stationary frame to rotating frame or (vice versa) can be easily done. The stationary frame is called α-β frame, on the other hand, the rotational frame is called d-q frame as Fig. 2.

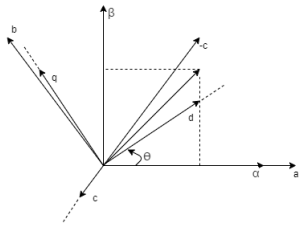


Fig. 2. α-β and d-q frame

A. Clarke & Inverse Clarke Transformation

Transforming from the 3-ϕ reference voltages or currents frame (a,b,c) to two-axis stationary frame (α,β) is called clarke or α-β transformation.

$$\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \frac{2}{3}\begin{bmatrix}1 & -\frac{1}{2} & -\frac{1}{2}\\ 0 & \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2}\end{bmatrix}\begin{bmatrix}a\\ b\\ c\end{bmatrix} \tag{1}$$

$$\alpha = \frac{2}{3}a - \frac{1}{3}b - \frac{1}{3}c \tag{2}$$

$$\beta = \frac{1}{\sqrt{3}}b - \frac{1}{\sqrt{3}}c \tag{3}$$

Using a + b + c = 0, the equations can be simplified as follows:

$$\alpha = a \tag{4}$$

$$\beta = \frac{a+2b}{\sqrt{3}} \tag{5}$$

Transforming from the two-axis stationary frame (α,β) to 3-ϕ voltages or currents frame (a,b,c) is called inverse clarke transformation.

$$\begin{bmatrix}a\\ b\\ c\end{bmatrix} = \begin{bmatrix}1 & 0\\ -\frac{1}{2} & \frac{\sqrt{3}}{2}\\ -\frac{1}{2} & -\frac{\sqrt{3}}{2}\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} \tag{6}$$

We can simplify matrix equation by using Eq. (4) and Eq. (5) as follows:

$$a = \alpha \tag{7}$$

$$b = -\frac{1}{2}\alpha + \frac{\sqrt{3}}{2}\beta \tag{8}$$

$$c = -\frac{1}{2}\alpha - \frac{\sqrt{3}}{2}\beta \tag{9}$$

B. Park & Inverse Park Transformation

Transforming from stationary frame (α,β) to rotating reference frame (d,q) is called park transformation.

$$\begin{aligned} d &= \alpha cos(\theta) + \beta sin(\theta) \\ q &= -\alpha sin(\theta) + \beta cos(\theta) \end{aligned} \tag{10}$$

Transforming from rotating reference (d,q) frame to stationary frame (α,β) is called park transformation.

$$\begin{aligned} \alpha &= dcos(\theta) - qsin(\theta) \\ \beta &= dsin(\theta) + qcos(\theta) \end{aligned} \tag{11}$$

C. Encoder Interface

There are four different designs to calculate θ and W values which are encoder, resolver, hall, and sensorless. The efficiency of calculation may vary according to motor types (PMSM, BLDC). While working BLDC motor type, the hall sensors provide better accuracy. On the other hand, while working PMSM motor type, the encoder sensors ensure better performance. In this study, we have chosen encoder structure to find θ and W values. The encoder structure is the quadrature encoder also known as incremental rotary encoder.

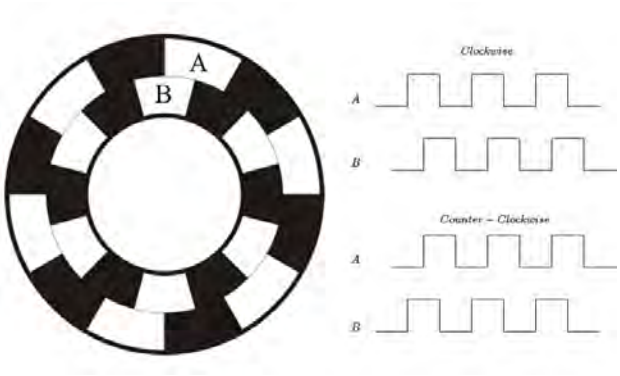


Fig. 3. Encoder Interface

As shown in Fig. 3, the direction of rotation can be easily determined. If the signal of B lagging to signal of A that means the direction of rotation is clockwise, otherwise counter-clockwise. Based on resolution of the encoder, the θ and the W can also be established.

D. PI Controller

The PI controller minimizes the error value based on input feedback and reference values. Feedback input stabilizes the unstable process due to the proportion process of PI. Since PI includes integration, PI controller output becomes an integral part of the given input. Implementation of the PI controller started with anti-windup integration, also referred to as integral windup. This feature gives the output accuracy of the PI Controller.

$$\begin{aligned} P_n &= K_p.E(n) \\ I_n &= K_i.T_s.E(n) + I_{n-1} \\ Y_n &= P_n + I_n \end{aligned} \tag{12}$$

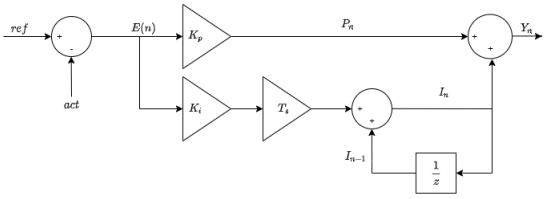


Fig. 4. PI Controller

E. Space Vector Modulation

Space Vector Modulation (SVM) is a sinusoidal wave generation technique that reduces Total Harmonic Distortion (THD) and can be used to increase the output voltage of the PWM drive. SVM has eight states that six active states, and two passive states. All of six states are driven by 3-ϕ two-level inverter. Thus, the motor has been driven. SVM is a technique that generates sine waves and feeds PWM. There's a lot of way to implement SVM. The min-max method has been used to perform SVM. Sampled voltages which has minimum value is called (V_{min}) and has maximum value is called (V_{max}). In order to calculate common voltage (V_{offset}) value as follows:

$$V_{offset} = -\frac{V_{max} - V_{min}}{2} \tag{13}$$

To skip 0-0-0 and 1-1-1 state because of increasing THD, the phase voltage value can be subtracted from common voltage value. By subtracting common voltage value to phase voltage value, it has achieved that to eliminate the third harmonic value of phase voltage.

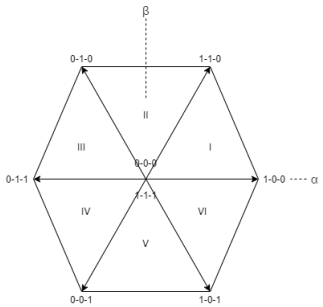


Fig. 5. SVM states

In Fig. 5, every state has max voltage value of $Vdc/\sqrt{3}$ where Vdc value is supply voltage value of 3-ϕ two-level inverter.

The theory behind the SVM is that finding V_{3H} which is third harmonic voltage and then subtracting V_{3H} from each phase voltage. The third harmonic voltage is formed as following:

$$V_{3H} = \frac{max(V_a, V_b, V_c) + min(V_a, V_b, V_c)}{2} \tag{14}$$

The calculated third harmonic voltage is subtracted to each phase voltage, space vector modulated reference voltages is found.

$$\begin{aligned} S_{V_a} &= V_a - V_{3H} \\ S_{V_b} &= V_b - V_{3H} \\ S_{V_c} &= V_c - V_{3H} \end{aligned} \tag{15}$$

III. Hardware Implementation & Results

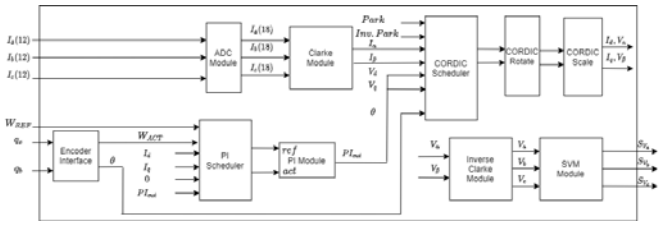


Fig. 6. Hardware Block Diagram of FOC

The Analog-to-Digital Converter (ADC) module has been prepared according to Xilinx ADC (XADC) wizard. XADC allows us to 12 bit precision, that's why we scaled up from 12 bit current values to 18 bit current values in this module. In order to construct this module, basic multiplier and simple state machine have been used. It has not been shown in the waveform for the accuracy of the test.

From the equations about clarke & inverse clarke transformation, low-level python script has been written first. Then, the result of the python script has been verified with given output vectors which are created by MATLAB. After that, Verilog module has been created. In order to be performed clarke & inverse clarke module, a simple multiplier, adder, and subtractor (MAS) module have been used.

In order to build encoder interface, the MATLAB simulation results have been taken first. Because the number of bits has been chosen as 18 for current precisions, the θ and the W precisions have been chosen same as current. The encoders have different resolution number. That's why, we have chosen a specific resolution number. According to resolution number, we have determined angle and speed factor values. The angle and speed factor values have been used to assign the θ and the W values by scaling 18-bit value. During the testbench of the FOC methodology, the encoder values have been determined, then they applied to the FOC methodology.

In order to implement park & inverse park transformation equations, the Coordinate Rotation Digital Computer (CORDIC) approach has been used. Firstly, the test vectors have been prepared by using MATLAB. Then, the low-level python script has been written according to the test vectors. After, the result of the python script has been verified with given output vectors. Following that, Verilog module has been created. While performing CORDIC algorithm, calculations have been made by applying fixed point theorem. Look Up Table (LUT)s and MAS module have been used instead of calculating the sine and cosine angle value for each time. Therefore, we have not only get rid of from the complexity of the calculation, but also from the area growth. For implementation of the park transformation, the design has been completed by taking only the reverse of the input angle (θ). Besides, the entire system has been scheduled to speed up our computations. The pipelining and resource sharing have been realized effectively.

For building PI controller module, the low-level python script has been prepared. Then, the output vectors which belong to PI module have been constructed by using MATLAB. Following that, the accuracy of the python script has been confirmed by comparing MATLAB vectors and python results. According to python results, the generic PI controller module have been generated and realized as I_d PI, I_q PI and Speed PI. In particular, the parameters of PI controller module should be initialized before running in motor control applications. For this reason, the PI controller module has been formed by using simple state machine. The K_i and K_p values have been chosen as 0.001525 and 0.097656 respectively.

The construction of SVM module, which is not very difficult theoretically, has been accomplished by the use of a simple state machine and MAS module.

For all the modules, the low-level python scripts which are 258 lines have been transformed to Verilog HDL code as 3258 lines with comment lines. The entire design have include MAS modules, CORDIC scheduler module, CORDIC rotate and CORDIC scale modules, clarke and inverse clarke modules, PI controller module, PI scheduler module, ADC module, encoder interface module, and SVM module.

The submodules which are clarke, inverse clarke, park, inverse park, ADC module, encoder interface module, PI controller and SVM module were put together to construct testbench of the FOC methodology. As shown in Fig. 1, speed PI controller is distinct from the clarke and park modules. That's why when the testbench is started, calculations of the speed PI controller can start. Because the PI controller module has initialization feature to keep a proper value before it is enabled, the PI controller output is the same as the PI controller's initial value. Keeping the PI controller's output value as its proper value avoids jerky design operation.

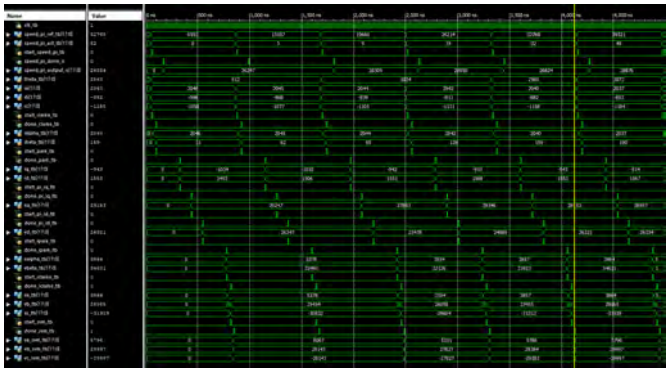


Fig. 7. Testbench Result of FOC Design

The entire design has 19 multiplication, 13 addition and subtraction, and 4 division. We have designed these operations

by using just four MAS module and binary shift logic. We have scheduled these operations by using resource sharing. Also, we have pipelined the ADC module, clarke module, and SVM module to reach higher throughput.

Based on the verification results, the ADC module have 5 cycles latency, the clarke and inverse clarke modules have 4 clock cycles latency, the park and inverse park modules have 23 clock cycles latency, the PI controller module has 11 clock cycles latency and the SVM module has 3 clock cycles latency. Encoder interface has 3 clock cycles latency but the delay of encoder interface is independent of system latency because of it is not dependent of any module.

The total latency of the FOC design is 84 clock cycles and the initiation interval of this design is one per 72 clock cycles. If the hardware clock frequency is 100 Mhz which is 10ns in terms of 1 clock cycle, the throughput value can be calculated as 1.39 Mbps. Also, maximum combinational path delay is 2.85ns and this delay is compatible with the design because of the hardware clock is higher than the maximum combinational path delay.

The comparison with papers that prefer FOC Methodology to control the 3-ϕ motor in their design is shown in Table I.

TABLE I
FOC IMPLEMENTATION COMPARISON

	FFs	LUTs	Sw. Freq.	Clk Freq.	FPGA
Kung	4174	15322	353 kHz	200 MHz	Altera Cyclone IV
Babu	5225	5514	N/A	200 MHz	Xilinx Virtex 5
Akin	1316	3172	400 kHz	50 MHz	Xilinx Spartan 3
Ours	1014	1245	1190 kHz	100 MHz	Xilinx Zynq-7020

In particular Xilinx and Altera different FPGA companies. Therefore, naming of logic units vary depending on which company you choose. There is a common index study which compares all FPGA companies in order to make them speak the same language in terms of Logic Block (LB)s [10]. Based on the common index, Altera LBs has been transformed into Xilinx LUTs as shown in Table I. Kung et al. have implemented their FOC design on Altera Cyclone IV [3]. Besides, they have implemented their design by using NIOS II processor. Akin [9] and Babu [5] have designed their FOC by using XSG. All this design approach may have caused the switching frequency to decrease. Based on Table I results, it can be said that the paper's method is superior to the other methods in terms of resource usage and maximum switching frequency.

IV. CONCLUSION

There are many advantages of a high-frequency approach in three-phase motor drive applications. Higher motor efficiency, low-cost filter, lower torque ripple, and faster control response can be among these advantages. In this study, we have offered a structure that is as fast as possible during consuming the least power. The paper's submodules (adc, clarke, inverse clarke, park, inverse park, encoder interface, PI, SVM), and testbenches that are used to verify those submodules modules have been written by using Verilog HDL. Besides, the theoretical results of those testbench modules have been written by using Python. On the other hand, this paper's work has been compared to the designs that include FOC Methodology in literature. The result of the comparison is that hardware implementation of this thesis work is provided superiority over other structures that are generated by using High-Level Synthesis (HLS) tools and HDL in terms of area and maximum switching frequency.

As for future work, this module will be applied to MATLAB Co-Sim block. Based on Co-Sim results, it can be applied to FPGA In the Loop (FIL) and it can be observed in terms of power and time. After then, it can be applied to real-time 3-ϕ motor control systems.

ACKNOWLEDGMENT

This work was supported by a TÜBİTAK BIGG project (no: 2180770).

REFERENCES

[1] J. Xu, L. Gu, Z. Ye, S. Kargarrazi, and J. Rivas-Davila, "Cascode GaN/SiC Power Device for MHz Switching," in *Proc. Applied Power Electronics Conference and Exposition (APEC)*, pp. 2780–2785, 2019.

[2] K. Shirabe, M. Swamy, J. Kang, M. Hisatsune, Y. Wu, D. Kebort, and J. Honea, "Advantages of High Frequency PWM in AC Motor Drive Applications," in *Proc. IEEE Energy Conversion Congress and Exposition (ECCE)*, pp. 2977–2984, 2012.

[3] Y. Kung, Hoang Than, Y. Lin, and L. Huang, "FPGA Based Speed Controller Design for a Ceiling Fan Motor," in *Proc. Int. Future Energy Electronics Conference (IFEEC) and ECCE Asia*, pp. 30–34, 2017.

[4] S. Suneeta, R. Srinivasan, and R. Sagar, "FPGA Based Control Method for Three Phase BLDC Motor," *Int. Journal of Electrical and Computer Engineering*, vol. 6, pp. 1434–1440, 2016.

[5] N. Babu and K. Athul, "The Field Oriented Control of Induction Motor Using FPGA," *Global Journal of Pure and Applied Mathematics*, vol. 11, pp. 1157–1170, 2015.

[6] J. Eriksson and L. Hermansen, "Rapid Prototyping: Development and Evaluation of Field Oriented Control Using LabView FPGA," Master's thesis, Mäladalen University, Sweden, 2011.

[7] K. Li, P. Evans, and M. Johnson, "SiC/GaN Power Semiconductor Devices: A Theoretical Comparison and Experimental Evaluation Under Different Switching Conditions," *IET Electrical Systems in Transportation*, vol. 8, pp. 3–11, 2018.

[8] M. Marufuzzaman, M. Reaz, and M. Ali, "FPGA Implementation of an Intelligent Current dq PI Controller for FOC PMSM Drive," in *Proc. Int. Conference on Computer Applications and Industrial Electronics (ICCAIE)*, pp. 602–605, 2010.

[9] O. Akin and I. Alan, "The Use of FPGA in Field Oriented Control of an Induction Machine," *Turkish Journal of Electrical Engineering and Computer Science*, vol. 18, pp. 943–962, 2010.

[10] O. Šćekić, "FPGA Comparative Analysis," tech. rep., University of Belgrade, 2005.

Co-Embedding Additional Security Data and Obfuscating Low-Level FPGA Program Code

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

Myroslav Drozd
Department of Information Systems
Odessa National Polytechnic University
Odessa, Ukraine
myroslav.drozd@opu.ua

***Abstract*—The paper is devoted to the issues of equivalent transformation of the FPGA program code in order to protect the integrity of this code. The paper considers an approach in which a digital watermark is covertly embedding into the FPGA program code. This digital watermark contains the monitoring data needed to implement the code integrity monitoring procedure. As a result, the digital watermark is hidden in the program code and forms a single whole with the program code. The embedded digital watermark does not change the characteristics of the device and does not make changes to its operation. Due to this approach, the fact of performing integrity monitoring in relation to program code is not obvious to an external observer. At the same time, monitoring data is also hidden and inaccessible. The paper proposes a method that improves integrity monitoring by jointly perform two processes: embedding secret additional data into the FPGA program code and obfuscating this program code. Moreover, both of these processes are proposed to be performed using one common system of equivalent transformations for the elementary units of the FPGA program code. Improved integrity monitoring is seen in the fact that the probability of detecting a digital watermark in FPGA program code is reduced. The paper describes an experiment showing the advantage of the proposed method over existing methods. The paper also provides recommendations on the possibility of using the method in areas that are related to the task of monitoring the integrity for FPGA program code.**

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

I. INTRODUCTION

FPGA chips are widely used as an element base for computer and control systems. FPGAs are program-controlled devices, that is, the operation of this type of chips is determined using program code. The operation of such microcircuits can be changed at any stage of the life cycle of the system by making changes to their program code.

The structure of FPGA chips [1, 2] is a matrix of elementary programmable calculating and specialized units. Each of these units is configured to perform a specific function using binary (low-level) program code. The matrix units are connected to each other and to the external FPGA pins using a programmable switching matrix. The specific switching option of this matrix is also configured using low-level program code.

Unlike microprocessors and microcontrollers, the computing process in FPGA chips is concentrated not in special large calculating nodes, but is distributed in the space of the matrix of FPGA elementary units [3]. Also, a distinctive feature of FPGA is the parallel principle of organizing the computing process. These features are responsible for several of the FPGA advantages have over microprocessors and microcontrollers. The main advantage of FPGA is the ability to provide significantly higher computing performance than microprocessors [4, 5].

The features and advantages of FPGA chips determine their applications. Chips of this type are usually used in cases where it is necessary to simultaneously fulfill two conditions: 1) it is required to provide high computing performance (which microprocessors cannot provide); 2) it is necessary to be able to change the operation of the chip at certain stages of its life cycle. Traditional areas of FPGA use are telecommunications, military and space applications, hardware implementation of cryptographic algorithms, implementation of complex industrial process control systems [6].

Low-level FPGA program code is a complex of binary data that determines the operation of a device. In these conditions, ensuring the integrity of the program code is one of the main components of the security for systems built on the basis of these chips. FPGA chips are very often used in critical applications. They are often used in control systems for high-risk objects [7]. Under these conditions, malicious violation of the integrity of the FPGA program code can lead to negative consequences [8]. Therefore, the creation of efficient methods and tools for ensuring the integrity of FPGA program code is an important and relevant task.

II. LITERATURE REVIEW AND GOAL OF THE PAPER

There are several ways to ensure the integrity of the program code. The main ones [9, 10]: organizational restrictions and differentiation of access to the program code; physical access restriction; cryptographic protection of program code; operational monitoring of integrity [11, 12]. Integrity monitoring is the most common among these ways. This is due to the ability to flexibly combine integrity

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

monitoring with other methods of counteracting interfering with the program code. In addition, unlike other ways, integrity monitoring is characterized by full coverage of the life cycle of a system built on programmable components.

Most often, the program code integrity monitoring is implemented using the procedure of double calculation of the hash sum [13]. When preparing the program code for monitoring, a hash sum is calculated for the information object of the program code. This calculation is performed using a predetermined hash function [14]. This hash sum is marked as a reference hash sum to integrity monitoring of this information object. The reference hash sum is stored in such a way that the monitoring system has access to it at any time. At the moment of execution of the integrity check procedure, the hash sum the program code information object is recalculated. The newly calculated hash sum is compared with the reference hash sum. If these sums coincide, the program code is recognized as integral. Otherwise, it is considered that the integrity of the program code is violated. Such programming code cannot be used to configure the device.

In the described integrity monitoring scheme, the main problem is the storage of and access to the reference hash sum. Several approaches are used to store a reference hash sum. One of the most common approaches is storing the reference hash sum in memory together with the information object of the program code [15]. Disadvantages of this approach: access to the reference hash sum is open; this creates the possibility of falsifying the hash sum or fitting the program code to the correct hash sum; the openness of the hash sum reveals the fact that integrity monitoring is performed. Also known is the approach within which the monitoring data are included in the information object of the program code [16]. Analysis of the structure of an information object makes it possible to identify the presence of monitoring data in it. Because of this, this approach has most of the disadvantages of the previous approach. In addition, there is an approach that is to store a reference hash sum in a remote database [17]. In this case, at the time of the integrity check, the reference hash sum is requested from this database. The disadvantage of this approach is the impossibility of excluding massive information leakage from such a database. Encrypting the data in this database does not completely eliminate the problem and creates the potential for falsification of monitoring [18].

There is a very effective approach to storing monitoring data as a digital watermark [19-21]. The monitoring data is embedded in the program code in the form of a digital watermark, without changing the size of the program code and the operation of the device. After this embedding, the digital watermark forms a single whole with the program code. It is not possible to select the location of the watermark in the program code. Extraction of monitoring data is possible only if there is a special key (steganographic key [22]), which shows the localization of the digital watermark. Without the presence of a stego key, an external observer cannot make a decision on the presence or absence of monitoring data in the information object. This leads to the fact that integrity monitoring is also hidden from an external observer.

To ensure the specified properties of a digital watermark, it is embedded in the FPGA program code using equivalent transformations [23, 24]. Equivalent transformations are performed in the space of program code of the FPGA elementary calculating units – LUT (Look Up Table) units [25]. The LUT is a programmable calculating unit that typically has 4 to 8 inputs for different FPGA families. This unit performs the calculation of one logical function of the input variables. The LUT is configured to calculate a specific function using a 2^n-bit binary program code, where n is the number of inputs of the LUT.

Equivalent transformations consist of elementary actions, each of which is performed for a pair of series-connected LUTs. Within each elementary action, the program code of the first LUT of the pair is bitwise inverted. This inversion is compensated for by bit rearrangement of the second LUT unit of the pair. The specific type of compensating rearrangement depends on which input of the second LUT unit is connected to the first LUT unit of the pair. When the weight of the input is k, the rearrangement consists in the exchange of places for groups of sequentially located 2^k-1 bits of the program code. In Fig. 1 shows two pairs of LUTs having equivalent program codes. The left pair of units has codes $code_1$ and $code_2$, respectively. In the right pair of units: unit LUT_1 has code $I(code_1)$ – bitwise inversion of code1; unit LUT_2 has code $P(v, code_2)$ – compensating rearrangement of the bits of the binary program code code2, the specific form of which depends on the weight v of the input of the LUT_2 unit, which is connected to the output of the LUT_1 unit.

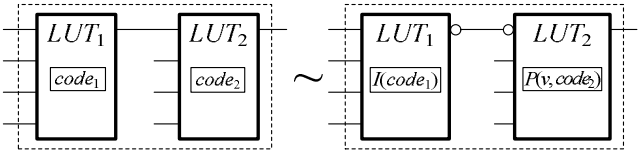


Fig. 1. Inversion and compensating rearrangement in the equivalent transformation for program code of LUT unit pair

As a result of the elementary action of the equivalent transformation, one bit of the digital watermark is embedded in the program code of the first LUT of the pair. In existing works, the only application for equivalent transformations of this kind in the task of ensuring integrity monitoring is proposed - embedding a digital watermark. However, these equivalent transformations have the potential to perform additional actions related to code integrity monitoring. We consider that such an additional action can be obfuscation of the FPGA program code, performed after the embedding of a digital watermark into it. The purpose of such obfuscation is to complicate steganalysis [26, 27] the information object of the program code, which is performed to detect a digital watermark in it.

Based on this, the *goal of this paper* is to make it difficult to detect the digital watermark embedded in the FPGA program code by applying the same equivalent transformations that are used in the embedding process.

III. Proposed Method for Joint Embedding of Digital Watermark and Obfuscation of FPGA Program Code

We propose a method that allows to perform two joint actions with the FPGA program code: 1) embed a digital watermark (which contains the data necessary for monitoring the integrity of the program code) into the program code; 2) reversibly obfuscate the resulting program code in order to complicate the steganalysis of this code. These actions are proposed to be implemented using a unified system of equivalent transformations, which in existing works is used only for embedding.

To describe the proposed method, we use the model of equivalent transformations (Fig. 2) of the FPGA LUT units program code. The model demonstrates the typical options for the links between LUT units in the process of equivalent transformations and the interaction between the program codes of these units.

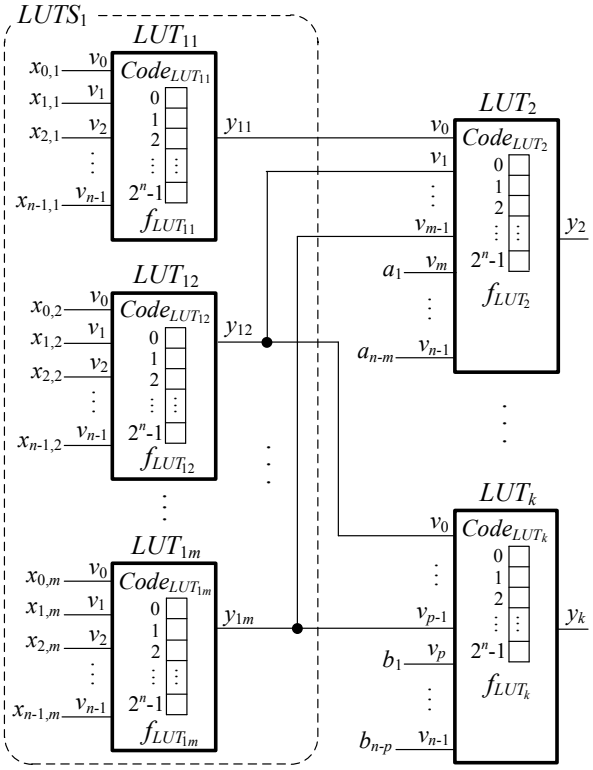


Fig. 2. Модель эквивалентных преобразований программных кодов блоков LUT в процессе внедрения и обфускации

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

The first principle of the method: the proposed method uses known equivalent transformations [23, 24]. In addition, the method, as well as the well-known embedding methods [28-30], requires the selection of a set of pairs $P_i = <LUT_{i1}, LUT_{i2}>$ of series-connected units from the set of LUT units.

The second principle of the method: the target place for embedding the digital watermark bit is LUT_{i1} – the first of the LUTs in pair of series-connected units.

The third principle of the method: the embedding of a digital watermark bit into the LUT_{i1} unit leads to an irreversible equivalent transformation of the P_i pair program code. Irreversibility is due to the fact that the procedure for performing the equivalent transformation of the unit program code depends on the ratio of the embedding bit and the target bit of the program code.

The fourth principle of the method: obfuscation of the program code of a pair P_i is carried out by a reversible equivalent transformation: 1) inversion for program code of one or more LUTs in the first level, the outputs of which are connected to the input of the LUT in the second level (Fig. 3); 2) compensating rearrangement for the bits of the LUT unit of the second level. Repeated this action leads to roll back the state of the unit program code. Due to this, the specified transformation is reversible.

The fifth principle of the method: the stego-key containing the embedding parameters and required to perform extraction has two types of parameters: parameters defining embedding and parameters defining obfuscation.

The sixth principle of the method: the parameters for the equivalent transformation of obfuscation are: 1) the set of the second LUTs LUT_{i2} in pairs P_i, for which the obfuscation was performed; This set is specified using a formal rule that generates the LUT_{i2} unit numbers. 2) the weights of the inputs of LUT_{i2} units from this set, which determine the type of bit rearrangement for the LUT_{i2} units. These weights can be determined in a fixed, pseudo-random, iterative or template way.

The seventh principle of the method: obfuscation of the program code is performed after embedding the bits of the monitoring digital watermark into this program code. To extract a digital watermark, first deobfuscation is performed, and then extraction is performed using known methods.

The proposed method is a sequence of stages, the implementation of which leads to the embedding of a digital watermark into the program code, as well as to obfuscation of this program code.

Stage 1. A stego key is formed, consisting of two components:

$$SKey = <K_{Emb}, K_{obf}>$$

where K_{Emb} is the data required for direct embedding and extraction of the digital watermark; K_{obf} – data required to perform obfuscation during the embedding stage and deobfuscation during the digital watermark extraction stage.

Stage 2. Embedding of a digital watermark into the FPGA program code is performed using one of the well-known embedding methods [28-30].

Stage 3. The value of the K_{obf} component of the stego key is specified:

$$K_{obf} = <LUTRule, InputsRule>$$

where *LUTRule* is a formal rule that allows to create a list of $LUTList_{obf} = <l_1, l_2, \ldots, l_n>$ LUT units, to the program code of which obfuscation is applied. It is further considered that the $LUTList_{obf}$ list components are the second LUT_{i2} units in the selected pairs of series-connected units (Fig. 2); *InputsRule* is a formal rule that allows to form a list $WList_{obf} = <Wl_1, Wl_2, \ldots, Wl_n>$, whose components $Wl_i = <wLUTi_1, \ldots, wLUTi_k>$ specify the set of weights for the inputs of the $l_i \in LUTList_{obf}$ units that are connected to the outputs of the invertible LUT_{i1} units of the pairs P_i.

The *LUTRule* specifies the order in which LUTs are listed to obtain a set of $LUTList_{obf}$. This rule can be specified in a fixed, pseudo-random iterative, or template way. The *InputsRule* specifies the weights for the inputs of the LUTs that participate in inversion compensation. This rule can also be described in a fixed, pseudo-random iterative, or template way.

Stage 4. For each $l_1 \in LUTList_{obf}$ unit, based on the corresponding $Wl_i \in WList_{obf}$ component, a list of $InvList_i = <invLUT_{1i}, \ldots, invLUT_{si}>$ of LUT units is formed (the program code of these units must be inverted during obfuscation).

Stage 5. Obfuscation of the program code is performed. For this, the following actions are performed for each of the LUT units $l_1 \in LUTList_{obf}$: a) the program code of each of the units $invLUT_{qi} \in InvList_i$ is bitwise inverted; b) the bits of the program code of each unit $l_1 \in LUTList_{obf}$ are rearrangement in accordance with the rearrangement rules defined for the set of weights $WList_{obf}$.

At the stage of digital watermark extraction, the stego key is considered known to the extraction party. In the case of using the proposed embedding method, the process of extracting a digital watermark from the FPGA program code consists of the following stages.

Stage 1. It is similar to *stage 4* of the embedding process and consists in obtaining lists of LUTs, the program code of which was inverted during the obfuscation stage.

Stage 2. Deobfuscation of the program code is performed. To do this, stage 5 of the digital watermark embedding process is repeated. Repeated obfuscation due to reversibility of transformations rolls back the program code to the state that the code had before obfuscation.

Stage 3. The digital watermark is extracted using the opposite method to that used in stage 2 of the embedding process.

IV. Experimental Assessment of the Proposed Method

The practical effect of using the proposed method is as follows. Embedding a digital watermark in the program code and obfuscation of this program code are performed using the same equivalent transformations. As a result of obfuscation, a massive change in the location of the bits of the program code of the LUT units is performed. This complicates steganalysis, the task of which is to decide on the presence or absence of additional hidden data in the program code. Also, due to the application of the proposed method, there is an additional obstacle to extracting the embedded digital watermark without knowing the stego key. This obstacle consists in the need to deobfuscate the program code before extracting a digital watermark from it.

An experiment was carried out to compare the proposed method with the existing approach. For this, software was developed that implements the method. The experiment involved eight FPGA projects of various sizes and design mission. Project synthesis was performed using CAD Intel Quartus [31]. FPGA Intel Cyclone IV [32, 33] were used as target chips.

During the experiment, in each of the experimental FPGA projects, a digital watermark was embedded using the existing, as well as the proposed method. Further, the program code of each of the projects was subjected to steganalysis. For this, several available software products were used [34-37]. It should be noted that digital watermarking and steganography technology for FPGA containers is now in its early stages of development. This technology has received significant development for multimedia containers (digital images, video, digital sound). For this reason, only software tools for steganalysis of multimedia containers are currently available. During the experiment, tools of this kind were adapted and used for the task of analyzing FPGA containers.

As a result of using software for steganalysis, estimates for the probability of the presence of embedded secret data in the program code were obtained. These estimates are shown in averaged form in Fig. 3. The diagram for each FPGA project shows an estimate of the probability (P) of presence embedded data without obfuscation and with obfuscation.

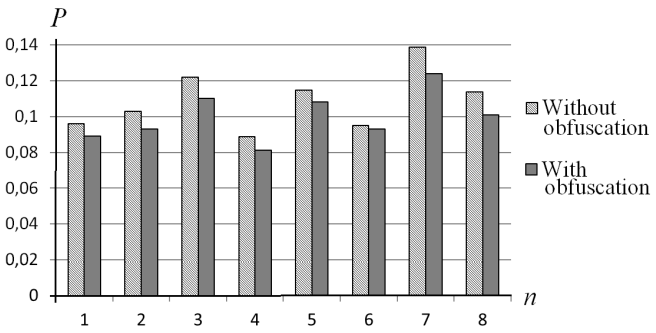


Fig. 3. Experiment results

For all eight projects, obfuscation of program code has reduced the estimate for the probability of presence embedded data. On average, this decrease was 10.32%. The resulting decrease in the probability P allows stating the effectiveness of the proposed method in terms of making a decision on the presence or absence of hidden data in the FPGA program code. The presence an additional deobfuscation stage characterizes an increase in the complexity of extracting a digital watermark without knowing the stego key.

V. Conclusions and Directions of the Further Research

The paper proposes a method that allows to combine the embedding of a digital watermark into the FPGA program code and the obfuscation process of this program code. An embedded digital watermark is used to store data required to monitoring the integrity of the program code. This approach makes the monitoring data hidden and does not reveal the fact of performing integrity monitoring in relation to the program code. The obfuscation of the program code (which is part of the method) complicates steganalysis and making a decision about the presence or absence of additional embedded data in the program code. Also, obfuscation further complicates the extraction of a digital watermark from the program code without knowing the stego key.

An experimental study of the proposed method has shown its effectiveness in comparison with the known approaches. The effectiveness of the proposed method is expressed in a decrease in the estimate (made using steganalysis) for the probability of the presence additional embedded data in the program code. The experiment showed a 10.32% decrease in this probability in comparison with the known approaches.

The areas of application of the proposed method are not limited to obfuscation performed after the embedding of a digital watermark. The method can also be used to reversibly hide the initial program code of an FPGA project. Also, the method can be applied for multi-level embedding of a digital watermark. At the same time, at some levels, embedding can be carried out together with obfuscation or replaced by obfuscation.

The experiment performed in this work was based on the use of software designed for classical steganalysis. This stegoanalysis is more designed to detect additional data in multimedia containers. Therefore, one of the important

directions for further research of the proposed method is its experimental evaluation using steganalysis software, which would be adapted for FPGA stego containers.

REFERENCES

[1] J. Andina, FPGAs: Fundamentals, Advanced Features, and Applications in Industrial Electronics. USA, Boca Raton: CRC Press, 2017.

[2] A. Raj, FPGA-Based Embedded System Developer's Guide. CRC Press, Boca Raton. USA, 2018. doi: 10.1201/9781315156200

[3] V. Sklyarov, I. Skliarova, A. Barkalov and L. Titarenko, Synthesis and Optimization of FPGA-Based Systems. Berlin: Springer, 2014.

[4] W. Vanderbauwhede and K. Benkrid, High-performance computing using FPGAs. USA, New-York: Springer, 2016.

[5] C. Unsalan and B. Tar, Digital System Design with FPGA. New-York, USA, McGraw-Hill, 2017.

[6] V. Hahanov, S. Chumachenko, E. Litvinova and M. Liubarskyi, "Qubit Description of the Functions and Structures for Computing," in Proc. of IEEE East-West Design and Test Symposium, Yerevan, pp. 88-93, 2016.

[7] 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.

[8] 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.

[9] 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.

[10] O. Drozd, V. Kharchenko, A. Rucinski et. al, "Development of Models in Resilient Computing," in: DESSERT 2019 – 10th IEEE International Conference on Dependable Systems, Services and Technologies, pp. 2-7. Leeds, UK, 2019. doi: 10.1109/DESSERT.2019.8770035

[11] W. Stallings, Cryptography and Network Security: Principles and Practice. 7th edn. United Kingdom, Harlow: Pearson Education Limited, 2017.

[12] M. Bishop, Computer Security. 2nd edn. USA, Boston: Addison-Wesley, 2018.

[13] J. Vacca, Computer and information security. 3nd edn. USA, Waltham: Morgan Kaufmann Publishers, 2017.

[14] 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.

[15] J. Katz, Digital signatures. Advances in Information Security. USA, New York: Springer, 2018.

[16] W. Conklin, et al. Principles of Computer Security, 4th edition. McGraw-Hill, 2015.

[17] W. Berchtold, M. Schafer and M. Steinebach, "Leakage detection and tracing for databases," in ACM Information Hiding and Multimedia Security Workshop, 2013.

[18] V. Hahanov, S. Chumachenko, W. Gharibi. and E. Litvinova, "Algebra-logical method for SoC embedded memory repair," in Proc. of The 15th International Conference Mixed Design of Integrated Circuits and Systems, MIXDES, pp. 481-486, 2008.

[19] F. Shih, Digital Watermarking and Steganography: Fundamentals and Techniques. 2nd edn. USA, Boca Raton: CRC Press, 2017.

[20] J. Fridrich, Steganography in Digital Media. USA, New York: Cambridge University Press, 2010.

[21] Ching-Nung Yang, Chia-chen Lin and Chin-chen Chang: Steganography and Watermarking. USA New York: Nova Science Publishers, 2013.

[22] M. Nematollahi, C. Vorakulpipat and H. Rosales, Digital Watermarking: Techniques and Trends. Springer, Singapore, 2017.

[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, pp. 654–661, 2017.

[25] H. Amano, Principles and Structures of FPGAs, Singapore: Springer, 2018. doi: 10.1007/978-981-13-0824-6

[26] A. Yahya, Steganography Techniques for Digital Images. Springer, 2018.

[27] 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.

[28] K. Zashcholkin, O. Drozd, R. Shaporin, O. Ivanova, and Y. Sulima, "Increasing the Effective Volume of Digital Watermark Used in Monitoring the Program Code Integrity of FPGA-Based Systems," in 2019 IEEE East-West Design and Test Symposium, EWDTS, pp. 53-58, 2019. doi:10.1109/EWDTS.2019.8884477.

[29] K. Zashcholkin, O. Drozd, O. Ivanova, and P. Bykovyy, "Formation of the Interval Stego Key for the Digital Watermark Used in Integrity Monitoring of FPGA-Based Systems." CEUR Workshop Proceedings, vol. 2623, pp. 267-276, 2020.

[30] 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.

[31] Intel Quartus, https://www.intel.com/content/www/us/en/software/programmable/quartus-prime/overview.html, last accessed 2020/07/30.

[32] Intel Cyclone FPGA series, https://www.intel.com/content/www/us/en/products/programmable/cyclone-series.html, last accessed 2020/07/30.

[33] Intel FPGA Architecture, https://www.intel.com/content/dam/www/programmable/us/en /pdfs/literature/wp/wp-01003.pdf, last accessed 2020/07/30.

[34] McAfee Steganography Analysis Tool, https://www.mcafee.com/enterprise/en-en/downloads/free-tools/steganography.html, last accessed 2020/07/30.

[35] Stegdetect, https://github.com/redNixon/stegdetect, last accessed 2020/07/30.

[36] StegoVeritas, https://github.com/bannsec/stegoVeritas, last accessed 2020/07/30.

[37] Zsteg, https://github.com/zed-0xff/zsteg, last accessed 2020/07/30.

Exploiting EEG Signals for Eye Motion Tracking

R. Kovtun[1], S. Radchenko[1], A. Netreba[1], O. Sudakov[1], R. Natarov[1,2], Z. Dyka[2], I. Kabin[2] and P. Langendörfer[2,3]

[1]*Medical Radiophysics Department Taras Shevchenko National University of Kyiv,* Ukraine

[2]*IHP – Leibniz-Institut für innovative Mikroelektronik* Frankfurt (Oder), Germany

[3]*BTU Cottbus-Senftenberg* Cottbus, Germany

***Abstract*—Human eye tracking devices can help to investigate principles of processing visual information by humans. The attention focus movement during the gaze can be used for behavioural analysis of humans. In this work we describe our experimental system that we designed for synchronous recording of electroencephalographic signals, events of external tests and gaze direction. As external tests we used virtual cognitive tests. We investigated the possibility to exploit electroencephalographic signals for eye motion tracking. Our experimental system is a first step for the designing an automatic eye tracking system and can additionally be used as a laboratory equipment for teaching students.**

Keywords—eye tracking, electroencephalographic (EEG) signals, exploratory factor analysis, visual behavioural analysis

I. INTRODUCTION

Eye tracking and devices to detect the human's gaze direction are used to study points of selective concentration i.e. spatial and temporal focus of human attention, and for complex investigations of human behaviour [1], [2]. Some of these sophisticated studies relate to the particularities of the human gaze and visual behaviours [3]. Most eye tracking devices are used to study the visual system, in training systems, in psychology, cognitive linguistics, for evaluation of information perception or reading speed, etc. [4]-[6]. Human eye tracking technologies are now being combined with statistical methods of data processing and machine learning [7], [21] allowing to create:

- advanced driving assistance systems and humanoid visual perception in robotics [8], [9];
- systems for education and learning based on the individual characteristics of students [2], [10], [14].

Additionally, eye tracking techniques are used in medicine as communication tools for patients with specific medical/physiological conditions such as Rett syndrome or amyotrophic lateral sclerosis [11], [12].

A large number of eye tracking systems based on eye motion measurements are available. Some eye-trackers use measured mechanical motion of the eye markers (i.e. the eye markers offsets) for determining the gaze direction [13]. Other systems use optical tracking of cornea movement for that [2], [11], [12], [14]-[17]. Optical eye-tracking systems are used different tracking algorithms [18]-[20]. They can be combined with tools for analysis and classification of medical diagnostic data [21], [22]. Eye-tracking systems based on an analysis of electro myopotential distribution recorded as electrooculograms are discussed in [23].

In this research we:

- demonstrate the relationship between gaze point direction and characteristics of electroencephalographic (EEG) signals by factor analysis of the EEG data;
- investigate the possibility to use only EEG signals for eye tracking and evaluate it using an optical eye tracking system.

For demonstrating the applicability of EEG signals for determining the gaze point direction we:

- developed a special experimental system that allows to measure EEG signals in parallel to optical eye tracking as well as behaviour tracking using a virtual cognitive test [24];
- synchronized the measured data and cognitive test events applying our synchronization system based on sound events [25];
- improved and extended a previously developed method [25] for merging data from various sources to study human behavioural reactions.

The experimental system that we designed can not only be used with the virtual water maze test that we selected as a representative example for our investigations, but with many other cognitive tests [24]. We expect that our system can be extended and improved in the future so that it can be used for investigating the exchange of information between:

- human(s) and technical system(s);
- human(s) and training/teaching system(s);
- human(s) and system(s) for visual behavioural analysis.

The rest of the paper is structured as follows. In section II we describe the experimental system that we developed. In section III we discuss the applicability of EEG data for determining the gaze point i.e. we analyze EEG data and demonstrate a relation between gaze point direction and characteristics of EEG signals. We evaluate the results of the proposed approach using a blind testing i.e. in our experiments the experimenter does not have any knowledge about the gaze direction of the test person. The experimenter determines the gaze direction using the proposed approach and compares it with the original gaze direction known by the test person. The paper finishes with short conclusions.

II. Developed Experimental System

Fig. 1 shows our experimental system schematically.

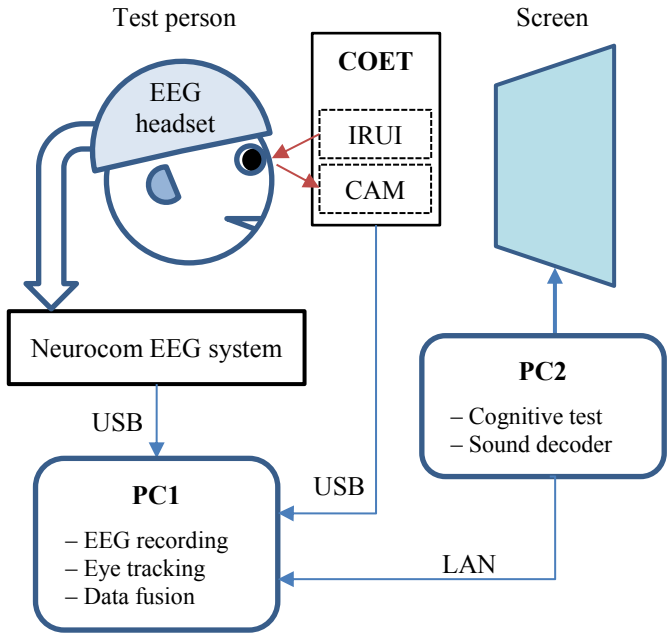


Fig. 1. Schematic representation of the developed experimental system.

In our experimental system we used two computers: PC1 and PC2. The data synchronisation between the computers is maintained by local-area network (LAN). PC1 receives data from the Neurocom EEG system and images from Contactless Optical Eyes Tracking (COET) developed by us. PC2 is used for running of the cognitive test – the virtual Morris maze test software [26], i.e. external events are generated in the maze test software. The external events are accompanied by specific sounds as we had no access to the Maze source code to register the events natively. The Maze test software uses files in a conventional sound format for playback, so we injected short sine wave sound at the beginning of each sound file. We used different frequencies for different events. The server-side of our custom-built sound decoding software, which also runs at the PC2, is able to easily capture sound from the Windows sound mixer during the walk through the Maze test, decode it in real time, and transmit corresponding events via LAN to the PC1. Computers are directly connected to each other, in order to minimize network latency. The overall delay between events and corresponding EEG signals is around 6 ms.

PC1 runs the Neurocom software to store EEG signals to the database. The client-part of our software receives events from the PC2 and writes them to the same database obeying the transmission delay.

For the eye tracking we used the third-party software solution 'EYE Writer' [12]. We calibrated this software to be able to use it with our particular camera. We tested this solution to verify the accuracy for the gaze detection. A high level of synchronization is achievable due to two additional electrocardiographic channels of the Neurocom system: we used one of them to write the event data, and the second one to store the gaze direction as a combination of two pulses of sine signals with different frequencies and durations.

Generally, the synchronization of the time and spatial information about the external events with other measured data is very important and a non-trivial data fusion task. The described data fusion technique allows to synchronously store not only events and gaze direction data with EEG data in one database, but also events from additional sources and other signals, as we did for the maze test software. The obtained data of the human behaviour should be suitable for measurement and further simultaneous analysis.

In general, our system operates as follows: the test person with a 19-chanel headset connected to the Neurocom EEG system also wears our COET device and looks at the computer screen with the Morris maze test and control the test. EEG signals with the events of the test and the gaze direction are synchronously recorded and stored to the database for the further analysis. Fig. 2 shows a test person wearing our experimental system.

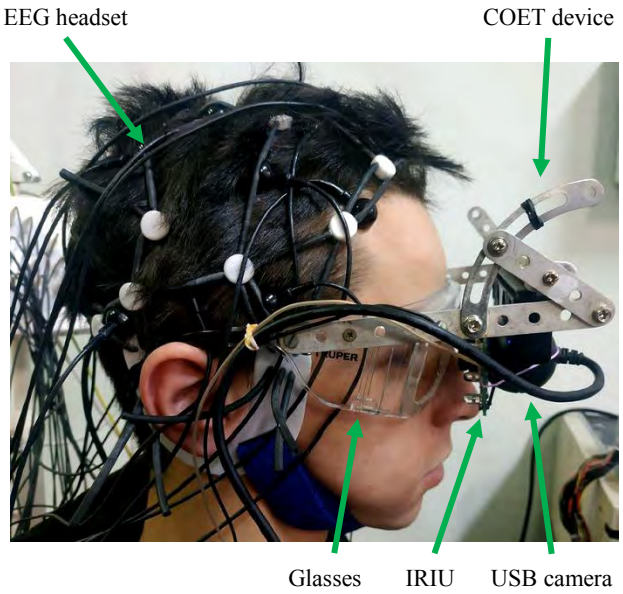


Fig. 2. Our experimental system on a test person.

Due to the inconvenience in simultaneously wearing the EEG headset and an eye-tracking device, determining the gaze direction using only EEG signals is by far more practicable, i.e. our approach can significantly decrease the complexity of evaluating and interpreting cognitive tests. This is the reason why we investigate the possibility to use only the EEG signals for determining the gaze point direction. In the rest of this section we explain some technical details of the EEG system used and the COET device developed.

A. The COET device

The COET device uses Contactless Optical Eye Tracking principles [12], [16]. The device is capable to detect the direction of eye movements and to track it. We used common low-cost components – plastic glasses, LEDs, metal frames etc., and a high-frequency commercial camera for designing the COET device.

The camera used is a USB PlayStation EYE camera with reasonable price, a sampling rate of 120 frames per second, and an eyeglass holder. This camera, like most portable colour USB cameras, has an infrared (IR) filter blocking infrared radiation needed for correct colour representation during photography. We replaced this filter by one fully rejecting visible light but transparent for IR spectrum in order to reduce the unwanted glare from the cornea during the detection of the eye movement.

Additionally we constructed the InfraRed Illumination Unit (IRIU) that includes 8 LEDs. The uniformity of the lighting influences the gaze point determination. We selected the number and location of LEDs due to their polar radiation pattern thus the IRIU uniformly illuminates the area around the eye cornea. The modified camera and the IRIU were mounted on the eyeglasses using metal frames. We performed some adjustments of the IRIU and camera positions to achieve the best illumination and visibility of the cornea.

B. EEG system

The EEG headset used in our experiments is a 19-channels electroencephalographic system Neurocom. This system is a very powerful Windows PC electroencephalographic system, designed for a wide range of neurological studies (EEG recording, visual and auditory evoked potentials, video EEG, neurofeedback etc.) [27]. The Neurocom System operates with EEG signals in the range from 1 μV to 12 mV. It has a sampling frequency of 500 Hz and an effective noise value of 0.5 μV in the frequency band from 0.15 Hz to 100 Hz that covers most of the EEG oscillation bands.

The system acquires a 24 bit digital signal per channel and stores 21 data components (19 EEG channels and two additional electrocardiographic (ECG) channels) in the local computer database. The EEG system communicates with a personal computer via USB. The system is capable to perform simultaneous common-mode noise reduction to over 120 dB for all 21 components of an analogue signal.

III. First Experiments Confirming The Concept

A. EEG Dataset preparation

EEG signals do not contain the direct information about the gaze direction. This information is “hidden” in the EEG signals. The shape of the measured EEG traces depends not only on the brain activity of the test person. The muscular activity of the test person can influence the shape significantly. This influence – artefacts – is a kind of noise and needs to be filtered or at least reduced before any analysis i.e. the measured EEG traces have to be prepared for the analysis. This is necessary to notice the neural processes information we need for our application.

During our experiments the test person sat and did not make any sudden movements. The main artefacts of muscular activity are caused by eyes blinking. Artefacts caused by unstable contacts of the EEG headset electrodes can be detected using the methods described in [28]. We removed the muscular activity artefacts applying the Independent Components Analysis (ICA). The ICA decomposition was performed using the approach introduced in [21] implemented in EEGLab v.15 software [29]. According to this approach the EEG data matrix X is presented as product of two matrices A and s. Matrix A contains new components that are statistically independent and orthogonal. Matrix s is a mixing matrix that contains contributions of EEG channels into matrix A. Eyes blinking and other miogram artefacts correspond to few independent components with large amplitude, low frequencies and large contributions from electrodes close to eyes end neck. These components are removed from matrix A and then EEG signals are reconstructed. Fig. 3 shows a part of the measured EEG traces with artefacts. Fig. 4 shows the same part of the EEG traces after being processed using ICA i.e. muscular activity artefacts were successfully removed. The EEG electrodes were located on the scalp of the test person corresponding to the international 10-20 System of Electrode Placement. All EEG channels are listed on the right side in Fig. 3 and Fig. 4.

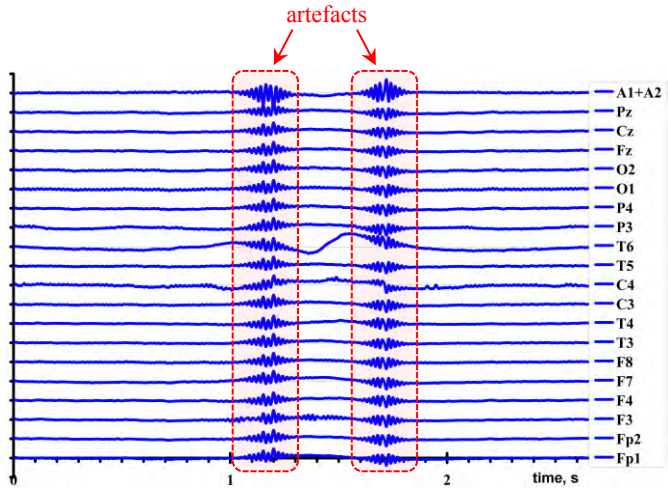


Fig. 3. Muscular activity artefacts in the measured EEG signal traces.

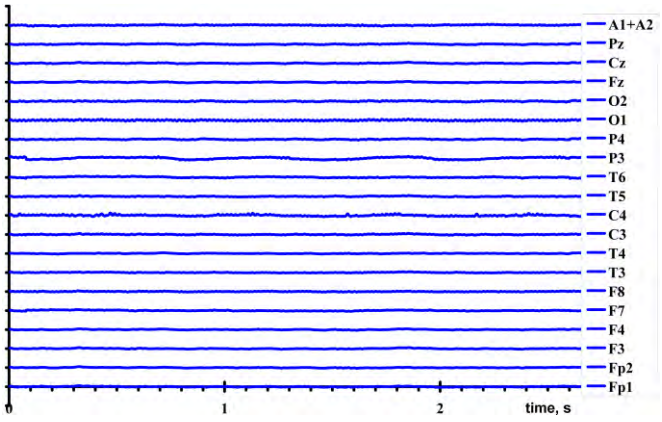


Fig. 4. Result of pre-processing the traces using PCA: muscular activity artefacts are successfully removed.

B. Electroencephalograms Factor Analysis

Our analysis of visual activity of a test person exploits the assumption that the EEG signals recorded synchronously with the gaze point should have common features. The models that connect EEG signals' features with gaze direction are unknown. In this work we approximated this relation using a linear model. EEG signals were processed by exploratory factor analysis to determine common components [30].

Factor analysis was performed for electroencephalograms corresponding to different positions of the gaze point on the screen. Our first step was measuring EEGs for training. For this purpose we used a specifically prepared image (Fig. 5), divided onto four equal quadrants. We recorded EEG data from three male test persons (subjects) about 20 years old. Each subject was asked to look at the center of each quadrant during a certain time (dots represents the point of gaze).

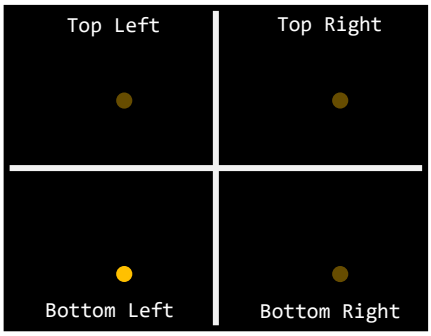


Fig. 5. Training image: test person looked at a centre of each quadrant in a predefined sequence.

We differentiate four diagonal gaze point directions (further also coordinates) only: top left, bottom left, top right and bottom right. Gaze point coordinates were determined using the COET device and the EYE Writer software. For each diagonal gaze point direction we measured 19 channel EEG signals during about 1 minute. We represented the 19 parallel measured signals as one signal i.e. we placed the signals from each channel serially after each other. Thus, we obtained four 19 minutes long traces. Each trace, i.e. the set of the feature vectors, corresponds to one gaze point direction. These sets of feature vectors for 3 test persons were analysed with IBM SPSS Statistics v.22 [31]. We applied Principal components analysis approach and Kaiser criterion. Fig. 6 shows the components' eigenvalues calculated for each test person.

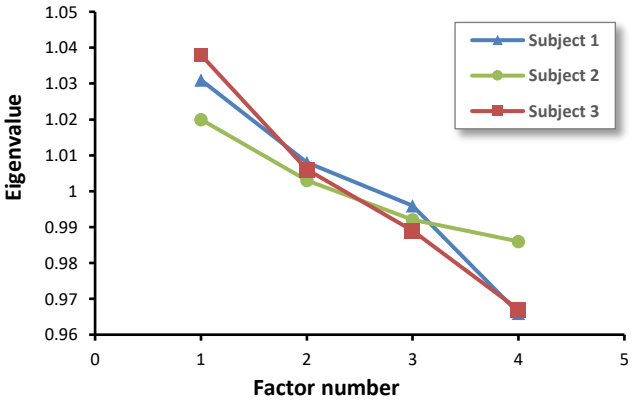


Fig. 6. Latent components eigenvalues for three test persons.

There are only two factors with eigenvalues higher than 1 (see Fig. 6). Thus, we concentrated on these two components only. The averaged scores of these components for different gaze point directions are shown in Fig. 7.

The data in Fig. 7 confirm our assumption that spatial and temporal dependences of EEG signals are associated with subjects' actions and different spots of human attention. The directions of gaze points are well defined by main components' eigenvalues. We used this fact for determining the "unknown" gaze point direction.

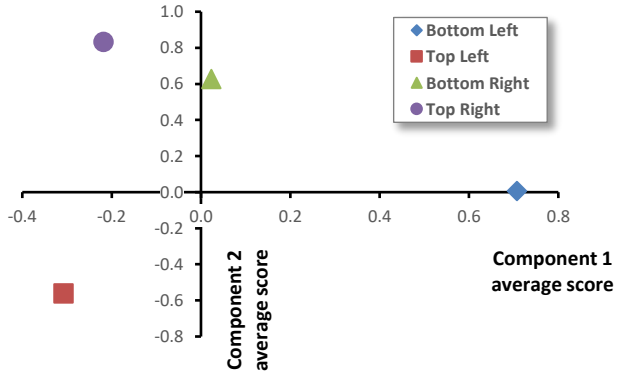


Fig. 7. Averaged scores of two main components for 4 gaze point directions.

C. *Blind test experiments*

To confirm the possibility to determine the gaze point direction by EEG data only we performed two experiments. Only one test person – Subject 3 – participated in these experiments. EEG signals were measured during 1 minute for each of two gaze point directions that were chosen by the test person himself. The first gaze direction was chosen to be identical to one of the indicated training directions (experiment 1). The second gaze direction was chosen different from the indicated training directions (experiment 2). It was expected, that values of components scores will be similar to the training data in case of identical gaze directions. The values of component scores obtained for the both experiments are shown in Table I.

TABLE I. COMPONENT SCORES CALCULATED FOR THE BLIND TEST EXPERIMENTS

Experiment	Component scores	
	for component 1	*for component 2*
1	0.014	0.682
2	0.475	–0.42

In experiment 1 the values of component scores are close to the component scores for the gaze point direction "bottom right". The test person as well as the eye tracking software confirmed the correctness of the obtained result.

The values of component scores calculated for the experiment 2 differ significantly from the data obtained for all 4 training gaze directions. Corresponding to information from the test person the chosen gaze point direction was "bottom" i.e. didn't coincide with any of training directions.

IV. CONCLUSION

Analysis of human behaviour using virtual cognitive tests is a basis for training and teaching systems. Determining the gaze point direction is a part of such systems. Synchronizing the events of the cognitive tests, gaze point direction of a test person and his/her reaction is a non-trivial task. In this work we

presented an experimental system that we developed for human behaviour analysis. Our experimental system allows to measure EEG signals in parallel to optical eye tracking as well as behaviour tracking.

While designing our experimental system we assumed that EEG signals can contain information about the gaze point direction. Determining the gaze direction using EEG signals only is by far more practicable: the complexity of interpreting cognitive tests can be significantly decreased using EEG signals i.e. without any eye tracking device and software for its synchronization with the EEG signals. We performed here only first experiments for confirming our assumption i.e. we demonstrated the relationship between gaze point direction and characteristics of electroencephalographic signals using factor analysis of the EEG data.

Accuracy and precision of determining the gaze point direction are important parameters. We will improve the experiments described here in our future work: the duration of a look at each gaze point for each test person has to be decreased whereas the number of gaze points for each test person as well as the number of test persons have to be increased significantly.

REFERENCES

[1] A. Ramirez Gomez and M. Lankes, "Towards Designing Diegetic Gaze in Games: The Use of Gaze Roles and Metaphors," Multimodal Technologies and Interaction, vol. 3, no. 4, p. 65, Sep. 2019.

[2] M. Q. Khan and S. Lee, "Gaze and Eye Tracking: Techniques and Applications in ADAS," Sensors, vol. 19, no. 24, p. 5540, Dec. 2019.

[3] B. Massé, S. Ba and R. Horaud, "Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 11, pp. 2711-2724, 1 Nov. 2018.

[4] Rudolf Netzel, Bettina Ohlhausen, Kuno Kurzhals, Robin Woods, Michael Burch & Daniel Weiskopf (2017) User performance and reading strategies for metro maps: An eye tracking study, Spatial Cognition & Computation, 17:1-2, pp. 39-64.

[5] Peter Kiefer, Ioannis Giannopoulos, Martin Raubal & Andrew Duchowski (2017) Eye tracking for spatial research: Cognition, computation, challenges, Spatial Cognition & Computation, 17:1-2, pp. 1-19.

[6] J. Rosch, J. Vogel-Walcutt. A review of eye-tracking applications as tools for training. Cognition, Technology & Work, vol. 15. no. 3, pp. 313-327, 2012.

[7] R. A. Naqvi, M. Arsalan, G. Batchuluun, H. S. Yoon, K. R. Park. Deep Learning-Based Gaze Detection System for Automobile Drivers Using a NIR Camera Sensor," Sensors, vol. 18, no. 2, p. 456, Feb. 2018.

[8] O. Palinko, F. Rea, G. Sandini and A. Sciutti. Eye gaze tracking for a humanoid robot, 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), Seoul, 2015, pp. 318-324.

[9] O. Palinko, F. Rea, G. Sandini and A. Sciutti: Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration, 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, 2016, pp. 5048-5054.

[10] EL Haddioui, Ismail & Khaldi, Mohamed. (2012). Learner Behaviour Analysis through Eye Tracking. International Journal of Computer Science Research and Application (IJCSRA) 2012-9564. 02. 11-18.

[11] R. Amantis, F. Corradi, A. Molteni, B. Massara, M. Orlandi, S. Federici, M. Belardinelli, M.L. Mele. (2011). Eye-tracking assistive technology: Is this effective for the developmental age? Evaluation of eye-tracking systems for children and adolescents with cerebral palsy. Assistive Technology Research Series. 29. 489-496..

[12] Open source eye-tracking system "Eye art" EyeWriter 0.20b: http://www.eyewriter.org/

[13] D. A. Robinson, "A Method of Measuring Eye Movemnent Using a Scieral Search Coil in a Magnetic Field," in IEEE Transactions on Bio-medical Electronics, vol. 10, no. 4, pp. 137-145, Oct. 1963.

[14] T. Pfeiffer, M. Latoschik, W. Ipke. Evaluation of Binocular Eye Trackers and Algorithms for 3D Gaze Interaction in Virtual Reality Environments. Journal of Virtual Reality and Broadcasting. Vol. 5, no. 16, p. 1660-1674, 2008.

[15] J. W. Lee, C. W. Cho, K. Y. Shin, E. C. Lee, K. R. Park. 3D gaze tracking method using purkinje images on eye optical model and Pupil. Opt. Lasers Eng. vol. 50, no. 5, pp. 736-751, May 2012.

[16] K. Murawski, T. Sondej, K. Rozanowski, O. Truszczynski, M. Macander, L. Macander. The contactless active optical sensor for vehicle driver fatigue detection. SENSORS, 2013 IEEE, Baltimore, MD, 2013, pp. 1-4.

[17] Wearable eye tracker https://www.tobiipro.com/product-listing/tobii-pro-glasses-3 28.02.2020

[18] S. Karthick, K. Madhav, K. Jayavidhi. A comparative study of different eye tracking system algorithms. AIP Conference Proceedings 2112, 020171 (2019).

[19] E Demjen, V Abosi, Z Tomori. Eye tracking using artificial neural networks for human computer interaction Physiol Res. 2011;60(5):841-4. doi: 10.33549/physiolres.932117.

[20] K. Krejtz, T. Szmidt, A. Duchowski, I. Krejtz. Entropy-based statistical analysis of eye movement transitions. Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14). Association for Computing Machinery, New York, NY, USA, 159–166.

[21] R. V. Byvalkevich, S. P. Radchenko, O. O. Sudakov. Tools for ultrasonic diagnostic image classification. Proc. IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2015, Warsaw, Poland, pp. 977-981.

[22] T. Frazier, E. Klingemier, M. Beukemann, L. Speer, L. Markowitz, S. Parikh, S. Wexberg, K. Giuliano, E. Schulte, C. Delahunty, V. Ahuja, C. Eng, M. Manos. Development of an Objective Autism Risk Index Using Remote Eye Tracking. Journal of the American Academy of Child & Adolescent Psychiatry, 2016, vol. 55, no. 4, pp. 301-309.

[23] F. Simini, A. Touya, A. Senatore, J. Pereira. Gaze tracker by electrooculography (EOG) on a head-band. 10th International Workshop on Biomedical Engineering, Kos, 2011, pp. 1-4.

[24] Assessment of the cognitive capabilities of humans and animals https://en.wikipedia.org/wiki/Cognitive_test

[25] O. Sudakov, G. Kriukova, R. Natarov, Distributed system for sampling and analysis of electroencephalograms. Proc. IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017, Bucharest, Romania, pp. 306-310.

[26] Morris, RG; Garrud, P; Rawlins, JN; O'Keefe, J (24 June 1982). "Place navigation impaired in rats with hippocampal lesions". Nature. 297 (5868): 681–3.

[27] EEG Systems NEUROCOM. Equipments catalog of XAI-MEDICA, Ukraine, Kharkiv. https://xai-medica.com/en/equipments.html 28.02.2020

[28] R. Natarov, O. Sudakov, Z. Dyka, I. Kabin, O. Maksymyuk, O. Iegorova, O. Krishtal, and P. Langendörfer "Resilience Aspects in Distributed Wireless Electroencephalographic Sampling," 9th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2020, pp. 1-7.

[29] Interactive toolbox for processing continuous and event-related EEG. https://sccn.ucsd.edu/eeglab/index.php 28.02.2020

[30] Anna B. Costello, Jason W. Osborne. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. Practical Assessment, Research & Evaluation. 7, vol. 10, July 2005 https://pareonline.net/pdf/v10n7.pdf 28.02.2020

[31] Powerful statistical software platform IBM SPSS Statistics: https://www.ibm.com/products/spss-statistics 28.02.2020

Metal Fillers as Potential Low Cost Countermeasure against Optical Fault Injection Attacks

Dmytro Petryk[1], Zoya Dyka[1], Jens Katzer[1] and Peter Langendörfer[1,2]

[1]IHP – Leibniz-Institut für innovative Mikroelektronik
Frankfurt (Oder), Germany

[2]BTU Cottbus-Senftenberg
Cottbus, Germany

***Abstract*—Physically accessible devices such as sensor nodes in Wireless Sensor Networks or "smart" devices in the Internet of Things have to be resistant to a broad spectrum of physical attacks, for example to Side Channel Analysis and to Fault Injection attacks. In this work we concentrate on the vulnerability of ASICs to precise optical Fault Injection attacks. Here we propose to use metal fillers as potential low-cost countermeasure that may be effective against a broad spectrum of physical attacks. In our future work we plan to evaluate different methods of metal fillers placement, to select an effective one and to integrate it as additional design rules into automated design flows.**

Keywords— optical Fault Injection attack; laser; reliability; security, countermeasure.

I. MOTIVATION

Wireless sensor networks (WSN) are more and more used in automation systems and in the area of critical infrastructure protection. One of the requirements for such devices is to keep the processed and transmitted data confidential and to ensure their integrity. This can be achieved by applying cryptographic algorithms.

The cryptographic strengths of a cipher algorithm depends according to the definition of Kerckhoff only on the used cryptographic key that is kept secret [1]. 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. The cryptographic keys have to be sufficiently long so that the time for brute forcing is sufficiently long. The situation changes dramatically if the attacker knows not only the input and output values but also intermediate values or physical parameters such as energy consumption and its distribution during the execution of the operation. Temperature, electromagnetic emission and other physically measureable parameters are a kind of "side effects". If the device is physically accessible the attacker can reveal the private/secret key analyzing side effects measured. These attacks are side channel analysis (SCA) attacks.

Another kind of powerful attacks are fault injection (FI) attacks. In these attacks faults are induced into an ASIC, e.g. in order to get access to internal data. FI attacks can be performed by various sources of faults: voltage, temperature, electromagnetic radiation, etc. The purpose of FI attacks is to inject a fault that switches the device into an erroneous operation mode. Exploitation of such an erroneous operation mode and monitoring its output may leak the secure data. In this work we discuss FI attacks performed with a laser, i.e. we concentrated on localized optical FI attacks. Optical FI attacks are feasible due to the internal photoelectric effect. This effect is based on interaction of silicon with laser light. Details about the internal photoelectric effect can be found in [2], [3].

Design and implementation of crypto hardware that is resilient against fault attacks is an extremely sophisticated task. At least, currently, there are no guidelines how to do it. The core idea discussed in this paper is to use metal fillers to prevent manipulation of devices by laser-based FI attacks.

The paper is structured as follows. Section II briefly describes the IHP technologies. Section III describes the optical FI setup used to perform the attacks described here. Section IV present a short description of the attacked chips and the obtained results. Section V discusses metal fillers as low cost countermeasure against physical attacks and compares it with existing radiation hardening techniques. Section VI concludes this work.

II. THE IHP CMOS TECHNOLOGIES

In order to prepare precise laser FI attacks knowledge about the internal structure of the chip is necessary. We use the IHP CMOS technology [4] as an example for our experiments. The IHP 8 inch wafer fab for research and production can manufacture chips in a 250 nm and in a 130 nm technology. In this section we give some details about these technologies. The knowledge of these details can be used not only by attackers for attack preparation but also by designers as effective countermeasures.

The thickness of the substrate of the 8 inch wafers is about 0.7 mm. It is usually thinned to a thickness of about 0.2-0.3 mm chips before they are used in devices. **Fig. 1** shows a cross-section of a chip in the IHP 250 nm CMOS technology schematically, i.e. wires in different metal layers and their interconnectors are zoomed in to illustrate the physical size of the chip structure.

Chips produced in the IHP 250 nm technology consist of 5 metal layers: 3 thin and 2 thick metal layers. The interconnectors

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

between metal layers are called vias. The bottom metal layer – metal layer 1 – is usually reserved for connecting transistors to power supply. In other metal layers the connection between gates is realized. The wires within a metal layer are (usually) parallel to each other while wires in neighboring layers are orthogonal to each other.

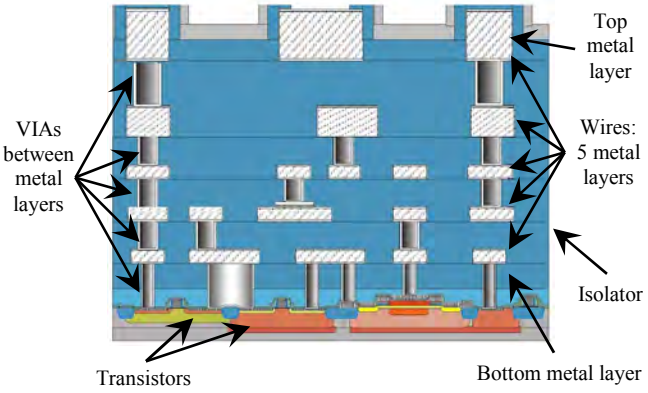


Fig. 1. Schematical cross-section of a chip on the example of the IHP 250 nm CMOS technology.

Chips produced in the IHP 130 nm technology consist of 7 metal layers: 5 thin and 2 thick metal layers. Due to technology requirements both IHP technologies, i.e. the 250 nm and the 130 nm technology have metal fillers.

Metal fillers are small metal structures – rectangles – placed in different metal layers. Metal fillers are applied as standard means to increase the mechanical stiffness of wafers during manufacturing process. If the global metal density of the metal layer does not meet the technology requirements it is filled with metal fillers. They are placed between the wires in each metal layer if required. The placement of the metal fillers is a mandatory step of the layout process that is performed automatically.

III. Setup for Optical FI Attacks at IHP

To perform laser based FI attacks we used a setup available at the IHP, see **Fig. 2**. It contains: a 1st generation Riscure Diode Laser Station (DLS) [5] placed in a safety box, a stable power supply, an oscilloscope and a PC with dedicated FI software.

The DLS consists of: a laser source, a spot size reducer, a microscope camera, a source of light for target illumination, a DLS body, an optical system and a high-precision X-Y positioning stage [6]. The DLS is equipped with two multi-mode laser sources. In 1st generation of the Riscure DLS only one laser source can be used simultaneously. According to [5] this DLS has following specifications:

- multi-mode laser sources:
  - red (808 nm), maximum power is 14 W;
  - infrared (1064 nm), maximum power is 20 W;
  - pulse duration in a range of 20 ns – 100 µs;
  - elliptical spot sizes of 60×14 µm², 15×3.5 µm², 6×1.5 µm² or 3×0.8 µm²;
  - filter: 0.1 %, 1 %, 10 %;
- single-mode laser source [7]:
  - red (808 nm), maximum power is 0.848 W;
  - pulse duration in a range of 2 ns – Continuous Wave (CW);
  - circular spot sizes of 15 µm², 4 µm², 1.5 µm² or 1 µm²;
- magnification objectives: 5×, 20×, 50×, 100×;
- X-Y table with 3 µm accuracy and 0.05 µm [5] step size.

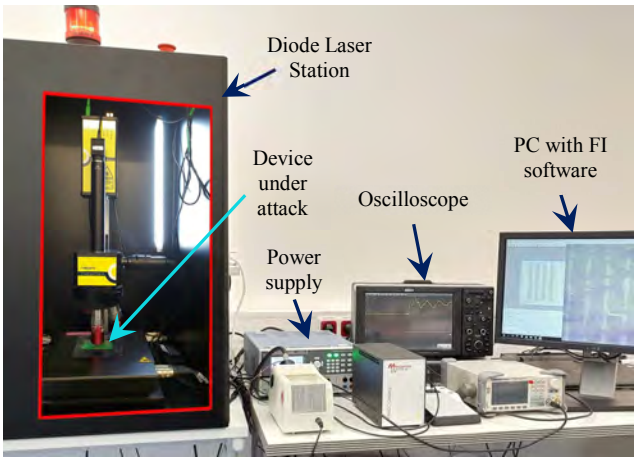


Fig. 2. Optical fault injection setup available at the IHP.

We applied the red multi-mode laser source in all our experiments described here. Thus all attacks have been performed through the front-side of the tested chips. Additional details about the optical FI setup can be found in [3].

IV. Effectiveness of Performed FI Attacks

A. Attacks against the IHP CMOS 250 nm technology

Our first device under attack (DUA) is IHP's "Libval025" chip manufactured in the IHP 250 nm technology with 5 metal layers. Originally, the chip was designed to measure signal propagation delays through different types of gates: invertors (INV), NAND gates, NOR gates and flip-flops (FF). Each libval-structure consists of 16 small independent circuits. Each circuit is a sequence of a single type of gates, e.g. a sequence of AND gates, or invertors only.

The structural scheme of Libval025 is shown in **Fig. 3**.

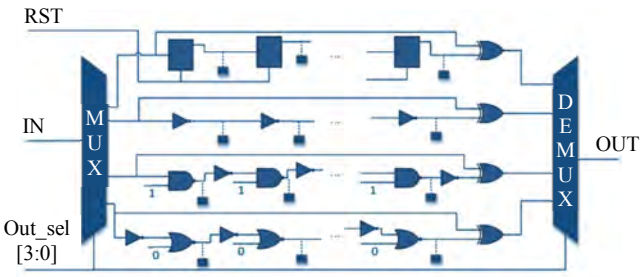


Fig. 3. Structural scheme of Libval025.

Fig. 4 shows Libval025 chips bonded on a PCB.

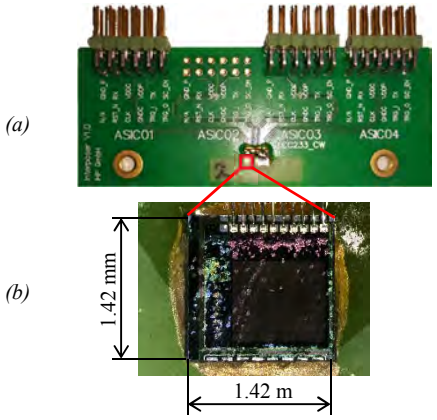


Fig. 4. The 3 bonded Libval025 chips on a PCB *(a)* and a single Libval025 chip zoomed in *(b)*.

We attacked all 3 Libval025 chips in our experiments.

The Libval025 chips described here were designed and produced about 20 years ago. We selected this chip for the experiments due to the fact that it was produced without metal fillers yet. Since Libval025 has no metal fillers the internal structure of the chip is clearly visible, e.g. through a microscope camera. **Fig. 5** shows the surface of a Libval025 chip without metal fillers captured using microscope camera with 100× magnification objective.

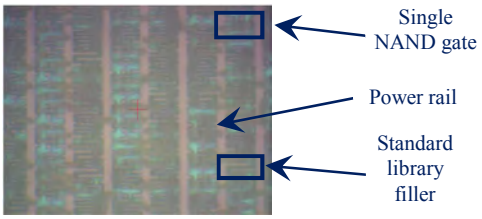


Fig. 5. Surface of a Libval025, captured using a 100× magnification objective.

In the Libval025 chips, single gates can be easily localized by optical inspection using a microscope camera or even a conventional microscope. Hence, optical FI attacks can be performed fast and effectively. Due to the parameters of our laser equipment and the size of the attacked gates, a single gate can be selected and attacked, i.e. we performed a localized FI attack. The single selected gate can be, e.g. a flip-flop of a register that can contain the secret/private key.

Results of front-side FI attacks on the Libval025 chips show that faults can be injected successfully in all 4 types of gates, i.e. FF, NAND, INV and NOR. The faults are repeatable for all 3 tested chips with a slight deviation of the applied laser beam parameters, i.e. intensity and/or pulse duration. Both transient and permanent faults were successfully injected. The latter was achieved by a significant increase of the laser beam power that subsequently led to the damage of the internal structure. A detailed description of the experiments done with the Libval025 chips can be found in [2]. A short summary of the attack results is given in TABLE I, see section *IV-D*.

B. Attacks against the IHP CMOS 130 nm technology

Next we attacked the Libval chip manufactured in the IHP 130 nm technology. We denote this chip as "Libval013" in the rest of this paper. The structure of Libval013 is the same as for Libval025, i.e. the chip contains 4 types of gates: INV, NOR, NAND and FF. Opposite to Libval025 the attacked Libval013 has metal fillers. They are placed in different metal layers. Due to the metal fillers the internal structure of Libval013 is not visible with a microscope camera. **Fig. 6** shows the surface of Libval013 captured by microscope camera using a 5× magnification objective and a part of Libval013 surface zoomed in using a 100× magnification objective.

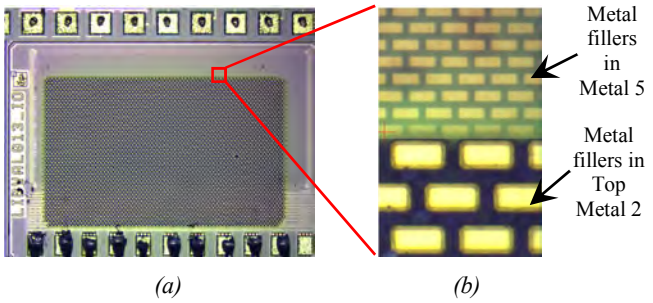


Fig. 6. Surface of a Libval013 captured using a 5× magnification objective *(a)* and a part of Libval013 surface zoomed in using a 100× magnification objective *(b)*.

The metal fillers in Libval013 have different sizes in different metal layers and can be placed on top of each other. The metal fillers in the Top Metal 2 are the biggest but the distance between the fillers – the "gap" – in this layer is also the biggest one (see **Fig. 6**–*(b)*). Thus we expected that in our experiments more successful FI will be observed when attacking the gates "placed" under the "big gaps" of large metal fillers.

The results of the front-side optical FI attacks on Libval013 confirmed our assumption. It was possible to inject faults in all type of gates. However the area of the chip that is sensitive to laser irradiation is reduced compared to Libval25. The state of the gates covered with metal fillers was not influenced in our laser experiments. Transient faults were successfully injected only in gates that are not covered with metal fillers. No damage of the internal structure was observed even if we illuminated the fillers with the maximum red laser beam output power over a relative long time (100 µs). The overall success rate of FI attacks is significantly reduced compared to Libval025.

C. Attacks against IHP RRAM structures

Additionally, we attacked the IHP Resistive Random Access Memory (RRAM) chips. The 4 kBit RRAM chips were manufactured in the IHP 250 nm technology. A single RRAM cell in IHP chips is based on a 1 transistor – 1 resistor (1T-1R) architecture. The memory element in the IHP RRAM cell is composed of a Metal Insulator Metal (MIM) stack. **Fig. 7** shows a Transmission Electron Microscope (TEM) image of an IHP RRAM cell based on the 1T-1R architecture.

The MIM structure is placed on top of Metal 2 and connected with Metal 3 through a tungsten via. The MIM structure is of interest here since we attacked a standalone RRAM chip, i.e. the NMOS transistor cannot be switched by laser irradiation.

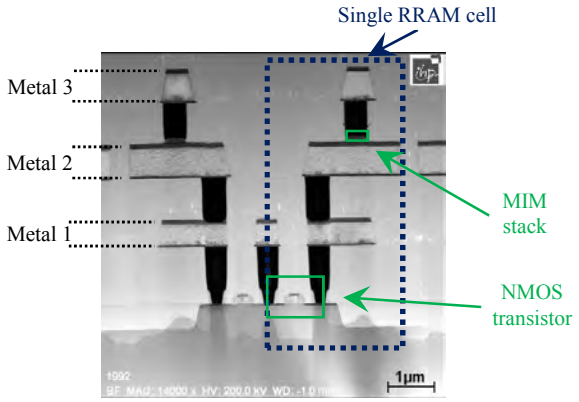


Fig. 7. TEM image of 1T-1R IHP RRAM cell, taken at IHP.

Additional details about IHP RRAM structures, i.e. switching behavior, purpose of transistor and MIM structure, can be found in [3], [8]. The IHP 4 kbit RRAM chips have metal fillers only in two metal layers, i.e. in Top Metal 2 and Top Metal 1 [9]. The size of all RRAM cells is the same. Thus, the placement of the cells as well as the one of metal fillers is periodical and the structure looks very regular. Due to the metal fillers the arrangement of the RRAM cells is not visible through the microscope camera. **Fig. 8** shows the attacked 4 kbit RRAM chip, the part of its surface was captured using a 100× magnification objective and a cross section image of the chip that was made with a Scanning Electron Microscope (SEM). The chip was prepared for the SEM-imaging using a Focus Ion Beam (FIB) cut in an IHP laboratory.

The front-side FI attacks on the RRAM chip show that it is possible to induce both transient and permanent faults into the chip. The latter is however achieved with significantly higher laser beam parameters than for Libval025. Analysis of obtained data shows that success of optical FI attacks depends on the location of metal fillers atop the RRAM cell. The metal fillers in the RRAM chip have different thickness but similar width and length. They are placed in Top Metal 1 and Top Metal 2, sometimes exactly under each other. Hence, areas that are not covered by metal fillers in Top Metal 2, can also not be covered by metal fillers in Top Metal 2. Thus, leaving "gaps" the laser beam can freely go through and illuminate the cell. In our experiments the RRAM cells placed under "gaps" were successfully influenced with the laser beam. Nevertheless some of the RRAM cells that are covered by metal fillers were also influenced but the success rate of the FI attacks for these cells is significantly lower. A detailed description of the experiments done with IHP RRAM chips and results of the FI attacks performed can be found in [3].

D. Attack results summary

TABLE I summarizes the results of the front-side optical FI attacks for different IHP chips.

The criterion of success of optical FI attacks was determined as follows (N is the percentage of successfully attacked gates/cells from all attacked gates/cells):

- very high: $(90 \le N \le 100)$ %;
- high: $(50 \le N < 90)$ %;

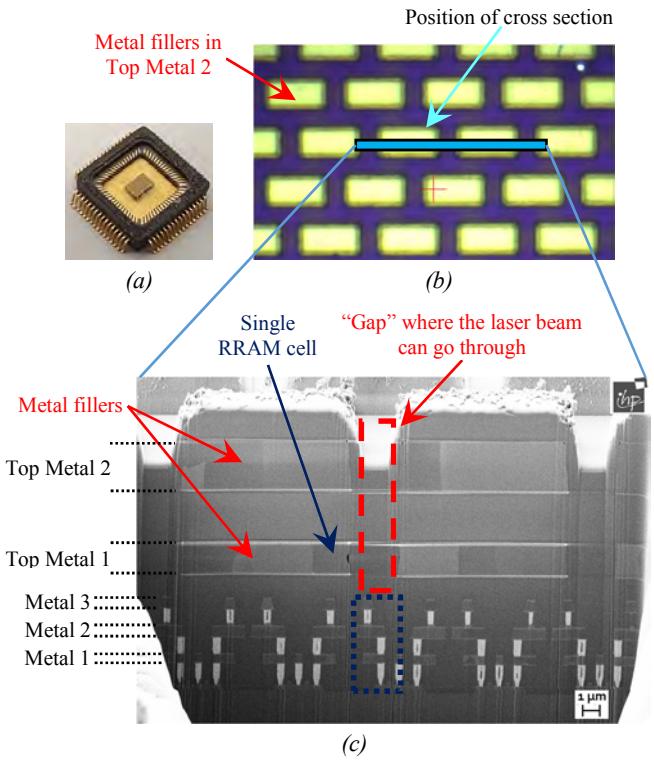


Fig. 8. Attacked RRAM chip *(a);* a part of the chip surface, captured using 100× magnification objective *(b)*; SEM cross section image (FIB cut) of the attacked IHP RRAM chip *(c)*.

TABLE I. SUCCESS OF OPTICAL FI ATTACKS FOR DIFFERENT IHP CHIPS

	Device under attack				
	Libval025	**Libval013**		**RRAM**	
Metal fillers, placement	no fillers	placed in Metal 5	placed in Metal 5 and Top Metal 1, and Top Metal 2	"gaps" between metal fillers	placed in Top Metal 1 and Top Metal 2
Success of FI attacks	very high	low	very low	moderate	low

- moderate: $(25 \le N < 50)$ %;
- low: $(10 \le N < 25)$ %;
- very low: $(0 \le N < 10)$ %.

V. METAL FILLERS AS LOW COST COUNTERMEASURE

Metal fillers and connecting wires are obstacles for laser beam propagation since they absorb/reflect the laser light making harder it for a laser beam to reach the attacked gate. The results of our experiments given in TABLE I confirm the fact that metal fillers reduce the success of front-side optical FI attacks significantly.

The idea to prevent optical FI attacks as well as localized electromagnetic analysis attacks using metal obstacles for a laser beam propagation is not new. For example in [12]-[14] it was proposed to use the metallization layers for supplying ASICs with VDD and GND as a countermeasure for semi-invasive

front-side attacks. Both supply voltages, implemented as metal planes, can be placed for example in top metal layers to prevent optical access to the transistor level while the device is fully functional. Additionally, it may be effective against microprobing, too.

Alternatively, efficient countermeasures can be designed to prevent/mitigate radiation influence using redundancy. These are hardware Triple Modular Redundancy (TMR) [10], Junction Isolated Common Gates (JICG) [11], Dual Interlocked Storage Cell (DICE) [12], etc. They have been studied intensely and already experimentally proved their fault-tolerance to radiation influence [10], [11], [12]. The resistance of the radiation hardened gates against manipulation as well their resistance against side channel analysis attacks have to be investigated. The need of this investigation is discussed in [15]. However implementation of such countermeasures usually requires hardware redundancy, e.g. triplication in TMR [10], doubling in JICG [11], doubling or triplication in DICE cells [12]. Due to this fact, such countermeasures require increased area on a silicon die compared to non-radiation hardened implementations. In order to increase the resistance of the redundant elements against laser fault injection the elements cannot be placed next to each other. For example, if TMR flip-flops will be placed close to each other a laser beam with a relatively large spot size, e.g. the size of two flip-flops, can influence them simultaneously with a high probability. Thus, placement of such redundant elements and arrangement of connection wires between them have to be considered. This usually leads not only to increased area but also to a modification of the automated standard design flow. On the other hand, since implementing these countermeasures requires additional active elements, the device's power consumption increases. Simultaneously, the increasing number of elements can cause a degradation of the performance. Thus, designers have to find a compromise between device performance, fault tolerance and resistance against SCA attacks.

Currently, metal fillers are a technology requirement of IHP that cannot be excluded from design flow. We propose to use the metal fillers as a kind of low-cost but effective countermeasure. To counter optical FI attacks efficiently, the placement of metal fillers has to be carefully evaluated. For example, default placement rules for metal fillers in the automatic design flow does not guarantee efficient mitigation against optical FI attacks since their coverage is not dense enough, i.e. a lot of cells are still not covered, e.g. in IHP RRAM chip. Thus, design rules with respect to metal fillers placement as well as their size and shape have to be reconsidered, i.e. a modification of the automated design flow is required. To comply with this task the areas that are sensitive to laser irradiation have to be determined for each type of gates. After it, the metal fillers can be placed so that all sensitive gate areas will be covered. Subsequently this "intelligent" placement of metal fillers methodology can be automated and implemented in the design flow.

In comparison to the countermeasures mentioned at the beginning of this section the metal fillers have several advantages. It is expected (but still has to be proven) that the intelligent placement of metal fillers will not cause a big overhead in chip area since it does not require doubling or triplication of elements, as metal fillers are more or less only arranged differently in the respective metal layers. The other advantage is that the metal fillers do not consume any power, i.e. the overall power consumption of the device does not increase. Thus, applying metal fillers and automatization of their intelligent placement can be a highly attractive, practical and low-cost countermeasure against optical inspection of chips, optical/laser FI attacks and – eventually even – localized electromagnetic analysis attacks. So, metal fillers can be a low-cost effective countermeasure against a broad spectrum of attacks. In our future work we plan to consider various methods of metal fillers placement in order to find the solution that successfully mitigates/prevents most of optical FI attacks.

In our future work we plan to consider various methods of metal fillers placement in order to find the solution that successfully mitigates/prevents most of optical FI attacks.

VI. Conclusion

In this work we discussed the results of our precise localized optical fault injection attacks. Our experimental results confirm the fact that the metal fillers placed in different metal layers of an ASIC can significantly influence the success of front-side optical FI attacks. Since the implementation of metal fillers is currently a requirement of IHP technology they can be used as a low cost countermeasure/mitigation technique if they cover the gate areas that are sensitive to optical FI attacks. In our future work we plan to investigate the sensitivity of the gates to laser irradiation. Our goal is to design a methodology for "intelligent" placement of metal fillers. The methodology has to be automated and integrated into the design flow.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 722325.

References

[1] Darrel Hankerson, Alfred Menezes, Scott Vanstone: Guide to Elliptic Curve Cryptography, Springer-Verlag New York, Inc., 2004, ISBN 0-387-95273-X

[2] D. Petryk, Z. Dyka and P. Langendörfer, "Sensitivity of Standard Library Cells to Optical Fault Injection Attacks in IHP 250 nm Technology," 2020 9th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2020, pp. 1-4.

[3] D. Petryk, Z. Dyka, E. Perez, M. Mahadevaiaha, I. Kabin, Ch. Wenger and P. Langendörfer, "Evaluation of the Sensitivity of RRAM Cells to Optical Fault Injection Attacks", accepted for Euromicro Conference on Digital System Design (DSD) 2020.

[4] IHP BiCMOS technology. URL: https://www.ihp-microelectronics.com/en/services/mpw-prototyping/sigec-bicmos-technologies.html

[5] Riscure. Diode Laser Station Datasheet, 2011. https://www.riscure.com/security-tools/inspector-hardware/

[6] Märzhäuser Wetzlar GmbH & Co. KG. Tango 2 desktop. https://www.marzhauser.com/de/produkte/steuerungen/tango-desktop.html

[7] Alphanov PDM laser source. URL: https://www.alphanov.com/en/products-services/pdm-laser-sources

[8] S. Dirkmann, J. Kaiser, C. Wenger and T. Mussenbrock, "Filament Growth and Resistive Switching in Hafnium Oxide Memristive Devices", ACS Applied Materials & Interfaces 2018 10 (17), pp. 14857-14868.

[9] F. E. Teply, D. Venkitachalam, R. Sorge, R. F. Scholz, H.-V. Heyer, M. Ullan, S. Dıez, and F. Faccio, "Radiation hardness evaluation of a 0.25

µm SiGe BiCMOS technology with LDMOS module", 2011 12th European Conference on Radiation and Its Effects on Components and Systems, Sevilla, 2011, pp. 881-888.

[10] V. Petrovic and M. Krstic, "Design Flow for Radhard TMR Flip-Flops", 2015 IEEE 18th International Symposium on Design and Diagnostics of Electronic Circuits & Systems, Belgrade, 2015, pp. 203-208.

[11] R. Sorge, J. Schmidt, C. Wipf, F. Reimer, R. Pliquett and T. Mausolf, "JICG MOS transistors for reduction of radiation effects in CMOS electronics", 2018 IEEE Topical Workshop on Internet of Space (TWIOS), Anaheim, CA, 2018, pp. 17-19.

[12] A. Krasnyuk and A. Kiseleva, "TMR vs. DICE schematic analysis", 2017 Internation Siberian Conference on Control and Communications (SIBCON), Astana, 2017, pp. 1-4.

[13] https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2014064275

[14] Z. Dyka, F. Vater, Ch. Wittke, A. Datsuk and P. Langendörfer, "Using Supply Voltage Metal Layers as Low Cost Means to Hinder Several Types of Physical Attacks", Proc. ISDF2014, 54 (2014).

[15] Z. Dyka, I. Kabin, D. Klann and P. Langendörfer, "Engineering of Resilient Crypto-Hardware: Balancing between FI and SCA Resistance", 22nd EUROMICRO Conference on Digital System Design and Software Engineering and Advanced Applications - Session on Work in Progress (Euromicro DSD & SEAA 2019), Kallithea, August 28 - 30, 2019, Greece.

Hyper Neural Network as the Diffeomorphic Domain for Short Code Soft Decision Beyound Belief Propagation Decoding Problem

Usatyuk Vasiliy, Egorov Sergey
South-West State University
Department of Computer Science, Kursk, Russia
Email:L@Lcrypto.com, sie58@mail.ru

***Abstract*—We proposed topological interpretation of the Tanner–Forney–Gross–Nachmani's Hyper Graph soft decoders based on Sourlas's Spin Glass reduction and Mezard's Replica Symmetry Breaking. Using it, we demonstrated reasons for uncertainty of the Neural Network loss function landscape and efficiency of replacing the $arctanh$ neural network activation function with the Nishimori Temperature $arctanh$ Taylor approximation. We compare the performance of short-length best known linear binary codes from Brouwer–Grassl codetable, Polar codes with sequence of frozen bits designed by Gaussian approximation and 5G eMBB Multi-Edge Type LDPC code with Base Graph 2 protograph on the AWGN-channel. The Sum-Product Flooding Scheduler decoder 50 iteration, Afterburn Saturated Min-Sum decoder, Ordered Statistics Decoder, Successive Cancellation Decoder with List size 32, Hyper Graph Neural Network under unfolding Belief Propagation decoder with Activation function Continues Metric Space relaxation according Nishimura temperature are used. The obtained simulation results are compared with the Finite-Length Polyanskiy theoretical boundary.**

***Index Terms*—Soft decision decoder; Multi-Edge Type LDPC; 5G; Polar Code; SCL; OSD; Afterburn Saturated Min-Sum**

I. Introduction

The development of machine learning has led to the emergence of new soft decision decoding methods, which current implementation allows decoding short-length codes. The question of comparing decoding methods using neural networks and state-of-the-art classical soft decision decoding methods arises.

Multi-edge Type (MET) QC-LDPC codes are the best error-correction code in LDPC codes family with the advantage of linear complexity soft decision decoding. These properties have become the reason for its widespread use for ultra high performance proprietor solution and standards: 5G eMBB [1], Deep Space Communication [2], TV physical layer standard ATSC 3.0 [3], fiber optic communication GPON [4], WDM Long-Haul [5], [6], and measurement matrix for Sub-Nyquist Sampling(Landau capacity reaching) in Compressed Sensing, [7]–[9]. 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 [10]. 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 [11]. Unfortunately at short lengths the suboptimal Belief-Propagation soft decision decoder of these codes suffers from trapping set and solitons, [12]–[18]. To solve this problem using idle resources of the decoder by modifying the scheduler the afterburn saturated min-sum decoder was proposed [19].

One of the most effective soft decision decoding methods for linear block codes was proposed by Fossorier as Ordered Statistics Decoding (OSD) method, [21]–[23]. This method is also applicable for improving LDPC Belief propagation decoding method and its approximations, [20]. Moreover Fossorier has shown that for OSD decoder with the order equal to $d_{min} \div 4$ under BPSK an AWGN channel achieves practically optimum Maximum Likelihood Decoding performance for a linear block code of minimum distance d_{min}.

Polar codes proposed by Arıkan is used in [1]. They achieve Shannon capacity of a binary-input discrete memory less channel using a successive cancellation decoder. The idea of constructing these codes and their corresponding successive cancellation decoder along with the concept Massey–Pinsker boosting the cutoff rate ([24], [25]), can be considered in the LDPC code decoder concept as for constructing a sequential scheduler over normal graphs that break short cycles and their corresponding trapping sets in order to obtain LDPC like codes. This idea is the development of the Forney's suggestion about Normal-graph cycles clustering proposed at [26], was independently obtained by Fossorier and Usatyuk published on the arxiv, [28] and represented to Dr. Arikan on Channel Coding Workshop, Moscow, Russia 2013, [29], [30]. The similar idea was published 5 years later by S. Cammerer, et al. [31]. One of the strongest method for breaking short cycles with linear complexity in the case of finite (non-asymptotic) code lengths is to use a list in successive cancellation decoding, [32]. Large list size allows similarly to a OSD decoder to achieve MLD perfomance. It is necessary to mention, that in this article we consider classical Arikan code's with Trifonov Gaussian Approximation frozen bit construction method, [33]. We do not consider Trifonov Polar Subcodes which greatly superior classic Arikan codes from code distance ([34]) and Arikan Polarization-adjusted Convolutional Codes which equivalent to Trapping set breaking using Zigangirov's

978-1-7281-9899-6/20$31.00

Spatially-Couple approach, [35].

The idea of unrolled iterative decoding algorithms was proposed by Weiberg [36] and was substantially generalized by Forney, [26], [27]. this approach has found wide application in Compressed Sensing and in the image denoising approximate message passing (D-AMP) algorithm and it modification, [8], [37], [38]. In [39] the similar idea was proposed to optimize parameters of linear code Min-Sum decoding method based on neural network optimization, in search of steady solution of non-linear system proposed by the famous Tanner paper [40]. The development of the idea of the smoothness of hyperparameters on a graph and the existence of attractors that make it possible to obtain an approximate in the variational sense solution of marginalization's the exponentially complex problem (or similar to this problem) using nonlinear optimization by a neural network led to the emergence of a wide class of hypergraph methods, [41]–[46]. Since the acceptable complexity is feasible only in the class of tree-like codes, topological properties linking the global structure of the graph and the local structure determine the characteristics of soft decoding dynamics. The problem of obtaining the variational approximation turns into a problem of dynamics at tree-like boundaries of a Margulis-Gromov hyperbolic space, [47], [48]. Fortunately, the possibility of choosing both a code and a decoder and its parameters probably allows us to hope for the solvability of this problem, at least for special non restriction on parallelism cases (like for Polar code), possible arising on certain lengths related to topology bundle. The smoothness of the nonlinear optimization space and the possibility of choosing the code/decoder make the directions of the hypergraph soft decision decoders the most promising area of finite-length capacity research.

Main contributions of the paper are topological formulation of the decoding problem, allowing to formulate the reason for the Neural Network loss function landscape uncertainty and a comparison of error-correction properties of soft-decision decoding methods under short length MET QC-LDPC based on 5G Base Graph 2 codes, Polar codes and the best Linear block codes known on such length.

II. Sum-Product soft decision method and it's Continues Space Statistical Physics Analog

A parity-check matrix specifying parity check equations of the code could be represented as a Tanner graph. For example the corresponding Tanner graph for parity-check matrix

$$H = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

is shown on Figure 1.

A bipartite graph can be generalized to a multi-graph. Multi-graph corresponding for parity-check matrix:

$$H_2 = \begin{pmatrix} I^1 + I^2 + I^7 & I^9 & I^{23} & 0 & 0 \\ I^{12} + I^{37} & I^{19} & 0 & I^{32} & I^{11} + I^{12} \\ 0 & 0 & I^{33} & 0 & 0 \end{pmatrix}$$

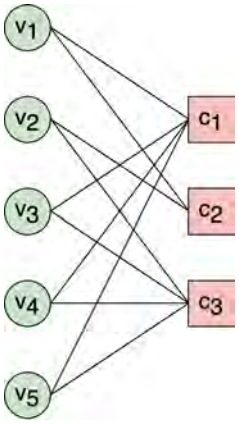


Fig. 1. Tanner Graph of H parity-check matrix

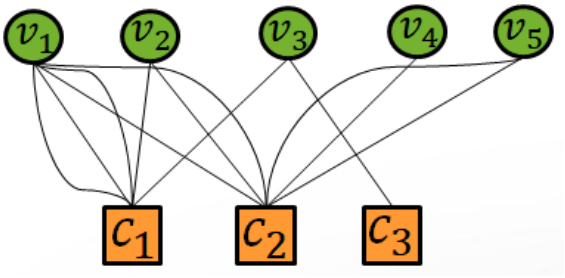


Fig. 2. Multi-graph of H_2 parity-check matrix

is shown on Figure 2, where I-circulant permutation matrix as at Multi-edge (Protograph) QC-LDPC codes, [10], [49].

Variable nodes v_i could be not only in prime field, but in Lie group, $(v_i \in GLG)$ or element from different algebra, [50]. Moreover multi-graph could be not bipartite and have multi-stage, example of such representaion Forney normal graph, [26]. Such generalization allow to consider non-linear code or equvivalent to them Tensor Networks, [27].

At the Sum-Product iterative method, known also as message passing method (Belief propagation, BP), log-likelihood ratios messages propagate from variable node (columns) to constrain node (check nodes, rows) and vice versa. The leftmost nodes $v_i, i = 1...n$ corresponds to a vector of the input channel log-likelihood ratios (LLR) $l \in \mathbb{R}^n$:

$$l_v = \log \frac{\Pr(c_v = 1|y_v)}{\Pr(c_v = 0|y_v)},$$

where $v \in [n]$ is an index of variable node and y_v is the channel output for the corresponding bit c_v, in which we want to correct the error.

One complete iteration of the BP method contains two computational steps. At the first half-iteration of BP method, when j is odd, a variable node computation is performed, in which the messages from variable, columns in Tanner graph are summed up:

$$x_e^j = x_{(c,v)}^j = l_v + \sum_{e' \in N(v) \setminus \{(c,v)\}} x_{e'}^{j-1}, \qquad (1)$$

where each variable node is indexed by Tanner's graph edge $e = (c, v)$ and $N(v) = \{(c, v)|H(c, v) = 1\}$, the set of all checks in which variable node v participates.

For even j half-iteration of BP method, the constrain (check) node performs the following computations:

$$x_e^j = x_{(c,v)}^j = 2arctanh\left(\prod_{e' \in N(c)\setminus\{(c,v)\}} tanh\left(\frac{x_{e'}^{j-1}}{2}\right)\right) \tag{2}$$

where $N(c) = \{(c, v)|H(c, v) = 1\}$ is the set of edges in the Tanner graph in which constrain node (check) c of the parity check matrix H participates.

The article [41] was suggested dual formulation of BP method. The dual formulation implies that the $tanh$ activation is moved to the variable node processing half iteration. In addition, a set of trainable hyperparameters weights are added. Hyperparameters w_e are shared across all half-iterations j of unfolded computation graph. If j is odd

$$x_e^j = x_{(c,v)}^j = \tanh\left(\frac{1}{2}\left(l_v + \sum_{e' \in N(v)\setminus\{(c,v)\}} w_{e'} x_{e'}^{j-1}\right)\right), \tag{3}$$

and if j is even:

$$x_e^j = x_{(c,v)}^j = 2arctanh\left(\prod_{e' \in N(c)\setminus\{(c,v)\}} x_{e'}^{j-1}\right) \tag{4}$$

The final stage marginalizes messages from the last constrain (check) node half iteration using the logistic activation function σ, and output n bits. The vth bit output at computation graph layer $2L + 1$, in the weighted version, is given by:

$$o_v = \sigma\left(l_v + \sum_{e' \in N(v)} \bar{w}_{e'} x_{e'}^{2L}\right), \tag{5}$$

where $\bar{w}_{e'}$ is a second set of learnable hypergraph parameter.

The neural network calculates the variational approximation to the correlation graph of the nodes of the corresponding parity-check matrix. Since the learning process requires a gradient, we pass to the corresponding continuous representation of the correlation tensor (matrix) - the Ising model, [27], [51], [57]. Consider Edward-Anderson Hamiltonian spin model:

$$H_{EA} = -\sum_{i=1,\ldots,n} \sum_{a=1,\ldots,m} C_{ij} J_{ij} \sigma_i \sigma_j,$$

where C_{ij} - connectivity matrix, the element of which are 1 if two spin interact and 0 otherwise, J_{ij}- is weigh interaction power between spins , σ_i- are Ising spins, n-number of column (variable nodes), m- number of rows (check nodes). The J_{ij}s give the strength of the two-spin interaction and are usually taken as independent random variables with a known probability distribution, where J_0 denote mean, $\triangle J^2$ the variance of distribution. H_{ECC} could be rewrites to infinite-range model:

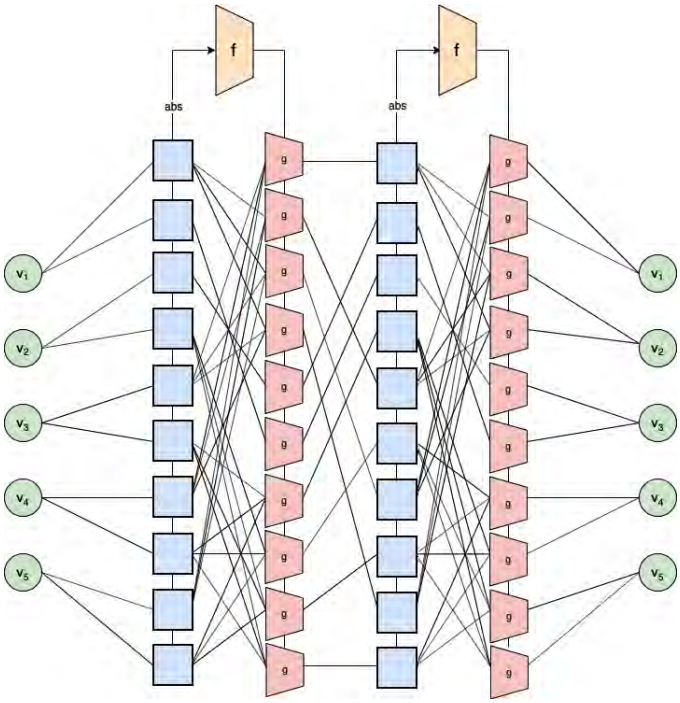


Fig. 3. Two iteration unfolding hyper graph neural network structures

$$H_{ECC} = -\sum_{p} \sum_{i_1,\ldots,i_p=1,\ldots,n} C_{i_1,\ldots,i_p}^{(p)} J_{i_1,\ldots,i_p} \sigma_{i_1} \ldots \sigma_{i_p},$$

where $J_0^{(p)}, \triangle\ J_{(p)}^2$ mean and variance of J_{ij}.

Let's a_i be n bits of codeword and $\sigma_i = 2a_i - 1$ the spins associated with them. The coded message, that is, the input to the channel, corresponds to the matrix element $J_{i_1,\ldots,i_p}^0 = \sigma_{i_1} \ldots \sigma_{i_p}$. Decoding and estimation of marginal corresponding to finding the ground state of Hamiltonian $H_{EA_{\inf}}$. The Js from channel equal to the J^0 plus noise. Code parity-check matrix is specified by the connective matrices C.

For one value of p in H_{ECC} and coordinate number $z_i = \sum_{j_2,\ldots,j_p} C_{i,j_2,\ldots,j_p} = z$ is independent of i, rate of the code equal $R = p/z$. If $p \to \infty$ we get random energy model, if $p = 2$ we get Sherrington-Kirkpatrick Spin Glasses for which each spin interact with all other spins. For decreasing compexity of ground state estimation in Sum-Product we can assume that each variable node has a tree "neighborhood. Such model of short range correlation at spin glass model used in Replica symmetry breaking method (Cavity method), [57]. Since the code which has the ability to correct error contain cycles, the residual curvature of the resulting universal covering tree contain curvatures. Such non-compensated curvature causes additional decoding errors.

Depending from noise power (Temperature) exist two phase: large temperature paramagneric phase (Symmetric phase, waterfall) when probability of error at any spin depends primarily on the generation from unfolded computation tree (large component in weight spectrum enumerator and linear size pseu-

docodewords) and low noise glass phase (symmetry breaking) when error depend from local curvature (code distance and sublinear size pseudocodewords), [52], [57]. The latter circumstance violates the codewords symmetry conditions, allows us to train this curvature compensation to different signal-to-noise values only in some variational sense, [40], [53]. Our ability to not sequential local compensation for curvature is determined by our tolerance to error. The use of a list, ordered statistics is inherently equivalent to trying to look into a neighboring area (closest voronoi regions) due to insufficient compensation for the curvature of space. In general case continuous and discrete curvature homotopy not equivalent without correct metric choice. The soft decoding problem with compensation of curvature is equivalent to the Topology complex thickening problem, [59], [60]. In essence, it is an in general ill-posed inverse problem, with a fundamental uncertainty due to the impossibility of taking the Poisson/Lie bracket, not the existence of heteroclinic solutions for sepatrice, [61], [62]. The correct choice of metric is known as matched metric property, [63], and depend on a priori knowledge or solution of metric learning problem, [64]. An important consequence of it is the phenomenon of loss function landscape uncertainty for the decoding problem and machine learning in general, [65].

III. Hyper Graph Neural Network Soft Decision Decoder

In paper [42], it was suggested Hyper Neural Network based metric learning, further adding learned components into the BP decoder. Specifically, they replace Eq. 3 (odd half iteration j) with the following equation:

$$x_e^j = x_{(c,v)}^j = g(l_v, x_{N(v,c)}^{j-1}, \theta_g^j), \tag{6}$$

where $x_{N(v,c)}^j$ is a vector of variable node degree length $d_v - 1$ that contains the elements of x^j that correspond to the indices $N(v) \backslash \{(c,v)\}$ of checks, and θ_g^j has the trainable hyper graph weights of neural network g at half-iteration j.

In order to make g adaptive to the current input messages reliability at every variable node, we employ a hyper network scheme and use a network f to determine its weights.

$$\theta_g^j = f(|x^{j-1}|, \theta_f) \tag{7}$$

where θ_f are the learned weights of network f. Note that g is fixed to all variable nodes at the same column. Scaled value parameters and offset parameters at unfolded BP computation graph allow to decorrelate variable nodes, partially compensate for curvature as it shown at papers [26], [40], [53].

Note that the messages x^{j-1} are passed to f in absolute value (Eq. 7). The architecture of both f and g does not contain bias (offset parameter, [39]) terms and employs the $tanh$ activations. The network g has p layers, i.e., $\theta_g = (W_1, ..., W_p)$, for some weight matrices W_i. The network f ends with p linear projections, each corresponding to one of the computation graph layers of network g.

For large temperature paramagneric phase (Symmetric phase, waterfall region), symmetry conditions is met, for such case sufficient to train f, g network under zero-codeword.

Another modification proposed in paper [42] is being done to the constrain (check) nodes in the Tanner graph. For constrain (check) nodes half iterations - even values of j, authors propose to use following equation, instead of Eq. 4.

$$x_e^j = x_{(c,v)}^j = 2 \sum_{m=0}^{q} \frac{1}{2m+1} \left(\prod_{e' \in N(c) \backslash \{(c,v)\}} x_{e'}^{j-1} \right)^{2m+1} \tag{8}$$

in which $arctanh$ is replaced with its Taylor approximation. The approximation is employed as a way to use finite temperature learning for which magnetization density function more flatter and closer to Nishimori Temperature $T = 1$, [56]–[58]. The Hyper parameters networks f and g have four layers with 128 neurons, two layer with 16 neurons respectively.

IV. Best Known linear block codes

The optimized linear block codes were taken from Brouwer–Grassl codetable, [54], [55]. Code with code length $n = 32$, information length $k = 20$ have $d_{min} = 6$ and weight enumerator: (0, 1), (6, 860), (8, 7402), (10, 43694), (12, 137497), (14, 259152), (16, 293445), (18, 201348), (20, 82818), (22, 19588), (24, 2608), (26, 158), (28, 5). Code with code length $n = 64$, information length $k = 20$ have $d_{min} = 19$ and weight enumerator: (0, 1), (19, 1049), (20, 2361), (23, 16643), (24, 28509), (27, 96211), (28,126813), (31, 202055), (32, 208898), (35, 157953), (36, 126827), (39, 45529), (40, 28591), (43, 4689), (44, 2239), (47, 157), (48, 49), (51, 2).

Code with code length $n = 128$, information length $k = 20$ have $d_{min} = 48$ and weight enumerator: (0, 1), (48, 16590), (56, 276048), (64, 583695), (72, 166320), (80, 5922).

Code with code length $n = 256$, information length $k = 20$ have $d_{min} = 48$ and weight enumerator: (0, 1), (112, 94860), (120, 198560), (128, 506175), (136, 175200), (144,73780).

OSD decoder was used from support material of Morelos-Zaragoza's book [66]. MET QC-LDPC codes were used from 5G eMBB standard Base Graph 2 protograph, except lower code rate which not supported, for which we extend parity-check matrix according code distance and EMD spectrum property, detailed described at paper [67]. Some of the source codes of the decoding methods given in the article are published by the authors in the public domain, [68].

V. Simulation Results

We have analyzed the performance of soft decision decoders under information length $k = 20$ MET QC-LDPC codes decoded by 50 iterations of BP decoder (flooding scheduler), Polar code constructed using Gaussian Approximation frozen bit sequence and decoded by Successive cancellation list 32 decoder, and the best linear codes with BPSK modulation in the AWGN channel by computer simulations. We simulate until 100 block errors are collected. Simulation results are represented in the Figure 4, Figure 5, Figure 6, Figure 7, codes length $n = 32, 64, 128, 256$ respectively. For comparison, also given finite-length capacity Polyanskiy Bound [69], [70]. From decoder result we see that BP decoder have more 1 dB gap on the block error rate 10^{-2} level and more than 2 dB gap

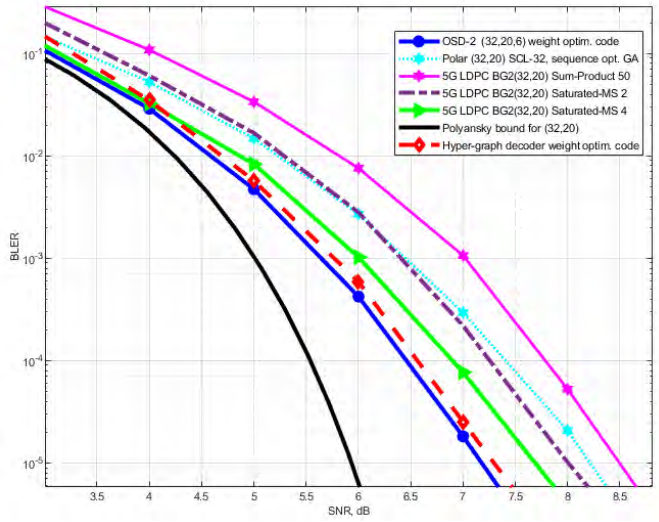


Fig. 4. Simulation result for Hypergraph Neural Network, OSD, Successive cancellation list Polar code, MET QC-LDPC codes under afterburn saturated min-sum decoders with code length $n = 32$ and information length $k = 20$

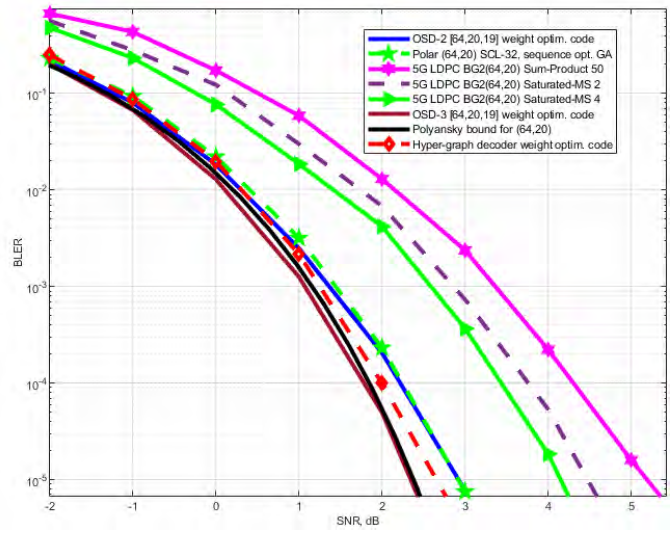


Fig. 5. Simulation result for Hypergraph Neural Network, OSD, Successive cancellation list Polar code, MET QC-LDPC codes under afterburn saturated min-sum decoders with code length $n = 64$ and information length $k = 20$

on the block error rate 10^{-5} level from finite-length capacity. OSD, Hyper Graph Decoder, Successive Cancellation Decoder with large List are the closest to Polyanskiy bound decoding results.

VI. Conclusion

Belief propagation decoder even using afterburn decoder still have gap from 1 to 2 dB gap on the block error rate 10^{-2} level to the finite-length capacity. The OSD and Hyper-graph decoder show best error correction performance on small pay load. Hyper Neural Network MAP decoder require matched metric property in order to guarantee continuous and discrete curvature homotopy equivalent.

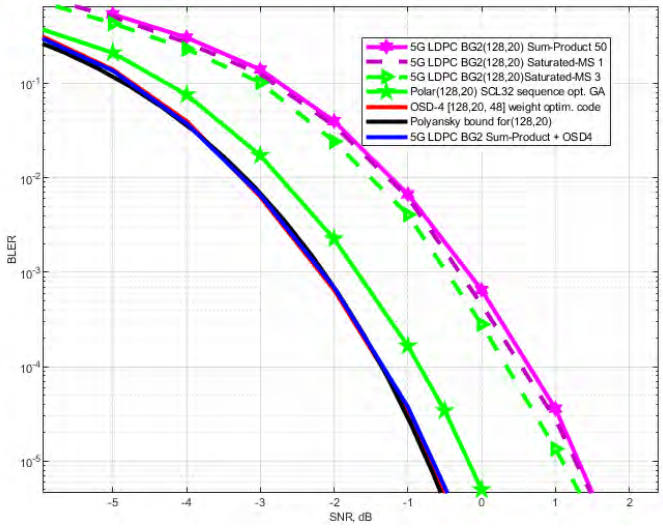


Fig. 6. Simulation result for OSD, Successive cancellation list Polar code, MET QC-LDPC codes under afterburn saturated min-sum decoders with code length $n = 128$ and information length $k = 20$

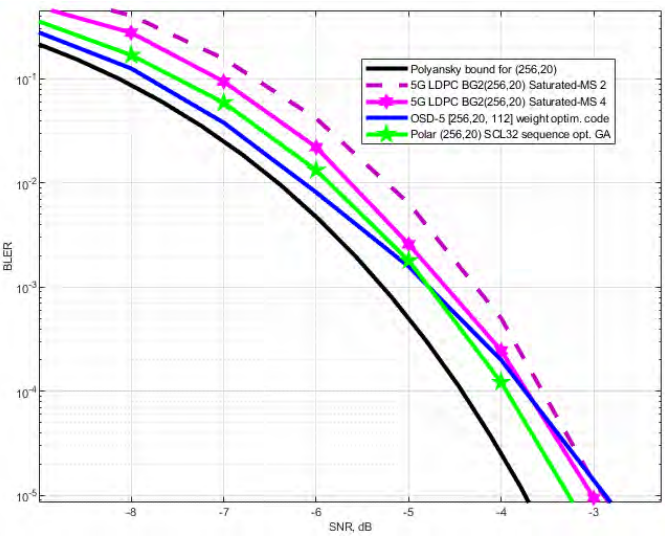


Fig. 7. Simulation result for OSD, Successive cancellation list Polar code, MET QC-LDPC codes under afterburn saturated min-sum decoders with code length $n = 256$ and information length $k = 20$

References

[1] 3GPP TS 38.212 V15.4.0: 'NR; Multiplexing and channel coding'

[2] CCSDS 131.0-B-3 Standard, Issue 3 September 2017 Washington, DC, USA, https://public.ccsds.org/Pubs/131x0b3e1.pdf

[3] K.-J. Kim et al., "Low-Density Parity-Check Codes for ATSC 3.0," IEEE Transactions on Broadcasting, 62(1), 189–196.

[4] IEEE P802.3ca 50G-EPON, May 23-26, 2017, New Orleans, LA, USA

[5] V. Usatyuk, I. Vorobyev, "Construction of High Performance Block and Convolutional Multi-Edge Type QC-LDPC codes," 42nd Intern. Conf. on Telecomm. and Signal Processing, 2019, pp. 158-163

[6] Patent US20190273511 - Usatyuk V.S. Generation of Spatially-Couple Quasi-cylic LDPC Codes, Public. Date 05.09.2019

[7] S. Jafarpour, W. Xu, B. Hassibi and R. Calderbank, "Efficient and Robust Compressed Sensing Using Optimized Expander Graphs," in IEEE Trans. on Inform. Theory, vol. 55, no. 9, pp. 4299-4308, 2009.

[8] D. Baron, S. Sarvotham, R. G. Baraniuk, "Bayesian Compressive Sensing via Belief Propagation," IEEE Trans. on Signal Processing vol. 58, no. 1, pp. 269-280, January 2010

[9] S. Pawar and K. Ramchandran, "FFAST: An Algorithm for Computing an Exactly k -Sparse DFT in $O(k \log k)$ Time," in IEEE Transactions on Information Theory, vol. 64, no. 1, pp. 429-450, Jan. 2018

[10] 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.

[11] J. Chenget et al., "Analysis of Circular Buffer Rate Matching for LTE Turbo Code," 68th Vehic. Techn. Confer., Calgary, BC, 2008, pp. 1-5.

[12] T. Richardson, "Error-floors of LDPC codes", Proc. 41st Annu. Allerton Conf., pp. 1426-1435, Oct. 2003.

[13] Cole C.A."Error floor analysis for an ensemble of easily implementable irregular (2048, 1024) LDPC codes," IEEE MILCOM, 2008, pp. 1-5

[14] X. Zheng, F. C. M. Lau, C. K. Tse and Y. He, "Construction of short-length LDPC codes with low error floor," 2008 IEEE Asia Pacific Conference on Circuits and Systems, Macao, 2008, pp. 1818-1821

[15] C. C. Wang et al "Finding all small error-prone substructures in LDPC codes", IEEE Trans. Inf. Theory, vol. 55, no. 5, pp. 1976-1999, 2009.

[16] X. Zhang, P. H. Siegel, "Efficient algorithms to find all small error-prone substructures in LDPC codes", Global Telecommun. Conf., pp. 1-6, 2011.

[17] K. Deka, A. Rajesh, P. K. Bora, "On the equivalence of the ACE and the EMD of a cycle for the ACE spectrum constrained LDPC codes," Intern. Symp. on Turbo Codes and Iterative Inform. Proc., 2014, pp. 67-71

[18] X. Zheng, F. C. M. Lau and C. K. Tse, "Constructing Short-Length Irregular LDPC Codes with Low Error Floor," in IEEE Transactions on Communications, vol. 58, no. 10, pp. 2823-2834, October 2010.

[19] S. Scholl, P. Schläfer and N. Wehn, "Saturated min-sum decoding: An "afterburner" for LDPC decoder hardware," Design, Automation Test in Europe Conference Exhibition (DATE), Dresden, 2016, pp. 1219-1224.

[20] M. P. C. Fossorier, "Iterative reliability-based decoding of low-density parity-check codes", J. Sel. Areas Comm., vol. 19, pp. 908-917, 2001.

[21] M. P. C. Fossorier, "Reliability-based soft-decision decoding with iterative information set reduction," in IEEE Transactions on Information Theory, vol. 48, no. 12, pp. 3101-3106, Dec. 2002,

[22] M. P. C. Fossorier, Shu Lin, "Soft decision decoding of linear block codes based on ordered statistics," IEEE ISIT. Theory, 1994, pp. 395-

[23] A. Valembois, M. Fossorier, "Box and match techniques applied to soft-decision decoding," Trans. on Inform. Theory, V.50(5), pp. 796-810, 2004.

[24] J. Massey, "Capacity, cutoff rate, and coding for a direct-detection optical channel," IEEE Trans.s on Comm., vol. 29, pp. 1615-1621, 1981.

[25] M. S. Pinsker, "On the complexity of decoding," Problemy Peredachi Informatsii, vol. 1, no. 1, pp. 84-86, 1965.

[26] G. D. Forney, "Codes on graphs: normal realizations," in IEEE Transactions on Information Theory, vol. 47, no. 2, pp. 520-548, Feb 2001

[27] G. D. Forney, "Codes on Graphs: Models for Elementary Algebraic Topology and Statistical Physics," in IEEE Transactions on Information Theory, vol. 64, no. 12, pp. 7465-7487, 2018

[28] M. Fossorier Polar Codes: Graph Representation and Duality, Dec 2013 arXiv:1312.0372v1

[29] Private communication with IEEE Fellow Dr. M. Fossorier. Congratulation to Marc briliant proof of scheduler interpretation of Polar code from LDPC point of view, sequential embedding to finite genus covering. 2013

[30] Usatyuk V.S. Improving of Polar decoding using LDPC cycle broking https://github.com/Lcrypto/Belief-Propagation-decoder-of-Polar-Codes

[31] S. Cammerer, M. Ebada, A. Elkelesh and S. ten Brink, "Sparse Graphs for Belief Propagation Decoding of Polar Codes," 2018 IEEE International Symposium on Information Theory, 2018, pp. 1465-1469

[32] I. Tal and A. Vardy, "List Decoding of Polar Codes," in IEEE Trans. on Inform. Theory, vol. 61, no. 5, pp. 2213-2226, May 2015

[33] P. Trifonov, "Efficient Design and Decoding of Polar Codes," in IEEE Trans. on Communications, vol. 60, no. 11, pp. 3221-3227, 2012

[34] Trifonov, P.; Miloslavskaya, V. Polar subcodes IEEE Journal on Selected Areas in Communications, 34(2):254-266 February 2016

[35] M Rowshan, A Burg, E Viterbo. Polarization-adjusted Convolutional (PAC) Codes: Fano Decoding vs List Decoding. arXiv:2002.06805, 2020

[36] Wiberg N. Codes and Decoding on General Graphs, Phd Thesis, 1996

[37] C. A. Metzler, A. Maleki, R. G. Baraniuk, "From denoising to compressed sensing," Trans. Inform. Theory, v.62(9), pp. 5117-5144, 2016.

[38] C. A. Metzler, Ali Mousavi, Richard G. Baraniuk, Learned D-AMP: principled neural network based compressive image recovery. 31st Intern. Conf. on Neural Inform. Proc. Systems., 1770–1781, 2017

[39] L. Lugosch and W. J. Gross, "Neural offset min-sum decoding," IEEE International Symposium on Information Theory (ISIT), Aachen, 2017, pp. 1361-1365. https://arxiv.org/abs/1701.05931

[40] Jinghu C. et al, "Improved min-sum decoding algorithms for irregular LDPC codes," Intern. Sympos. on Inform. Theory, 2005, pp. 449-453

[41] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein and Y. Be'ery, "Deep learning methods for improved decoding of linear codes," IEEE Journal of Sel.Top. in Sign. Proc., V. 12, no. 1, pp. 119-131, 2018.

[42] Nachmani E., Wolf L. Hyper-Graph-Network Decoders for Block Codes. Advances in Neural Inform. Proces. Systems, 32, 2019., pp. 2326–2336

[43] Tandler D. and Sebastian Dörner and Sebastian Cammerer and Stephan ten Brink On Recurrent Neural Networks for Sequence-based Processing in Communications,2019, http://arxiv.org/abs/1905.09983

[44] R. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec. GNNExplainer: Generating Explanations for Graph Neural Networks, Neural Information Processing Systems (NeurIPS), 2019.

[45] Naganand Y. M. et al. HyperGCN: A New Method For Training Graph Convolutional Networks on Hypergraphs, Advances in Neural Information Processing Systems (NeurIPS) 32, 2019, pp. 1509–1520

[46] Chami, I., Ying, R., Ré, C. and Leskovec, J. Hyperbolic Graph Convolutional Neural Networks. NIPS 2019.

[47] Margulis G.A. Discrete groups of motions of Nonpositive Curvature Manifolds. International Congress of Mathematicians, 1974, pp. 21-33

[48] M. Gromov Hyperbolic groups, in: Essays in Group Theory, S. M. Gersten ed., M.S.R.I. Publ. 8, Springer, pp. 75-265, 1987.

[49] D. Divsalar, C. Jones, S. Dolinar and J. Thorpe, "Protograph based LDPC codes with minimum distance linearly growing with block size," IEEE Global Telecomm. Conf., 2005., St. Louis, MO, 2005, pp. 5

[50] M. Korb and A. Blanksby, "Non-binary LDPC codes over finite division near rings," Intern. Conf. on Telecommunications (ICT), 2016, pp. 1-7

[51] Sourlas, N. Spin-glass models as error-correcting codes. Nature, 1989, 339, pp. 693–695

[52] Amraoui A., Montanari A., Urbanke R., How to find good finite-length codes: from art towards science. Eur. Trans. Telecomm.,2007, 18:491-508.

[53] Marc Vuffray M., Misra S., Lokhov A. Y., Chertkov M. Interaction screening: efficient and sample-optimal learning of ising models. 30th Intern. Conf. on Neural Inform. Proces. Systems, 2016, pp. 2603–2611.

[54] Andries E. Brouwer, Bounds on linear codes, in: Vera S. Pless and W. Cary Huffman (Eds.), Handbook of Coding Theory, pp. 295-461, 1998.

[55] Grassl, Markus. "Bounds on the minimum distance of linear codes and quantum codes." Online available at http://www.codetables.de.

[56] Sourlas, N.. (2007). Spin Glasses, Error-Correcting Codes and Finite-Temperature Decoding. EPL (Europhysics Letters). 25. 159.

[57] Mézard M., Montanari A. Information, Physics, and Computation Oxford Graduate Texts, 2009, 569 pages

[58] Baldock, R., Marzari, N. Bayesian Neural Networks at Finite Temperature. 2019, arXiv:1904.04154

[59] Adamaszek M, Adams H., Frick F. Metric reconstruction via optimal transport arXiv:1706.04876, 2018

[60] Niyogi P. et al, Finding the Homology of Submanifolds with High Confidence from Random Samples. Discrete Comput. Geom. 39, 1, 2008

[61] Kozlov V.V. Symmetries, Topology and Resonances in Hamiltonian Mechanics, 1996, 378 p.

[62] Boumal N. An introduction to optimization on smooth manifolds, Lie bracket, May 25, 2020 http://www.nicolasboumal.net/book

[63] M. Firer, J. L. Walker, "Matched metrics and channels", IEEE Trans. Inf. Theory, vol. 62, no. 3, pp. 1150-1156, Mar. 2016.

[64] Suárez J.L. et al, A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms and Software. arXiv:1812.05944, 2018.

[65] Li H. et al Visualizing the Loss Landscape of Neural Nets. NIPS, 2018.

[66] Morelos-Zaragoza R.H. The Art of Error Correcting Coding. Second Edition, John Wiley Sons, 2006 http://the-art-of-ecc.com/

[67] V. Usatyuk, S. Egorov, G. Svistunov, "Construction of Length and Rate Adaptive MET QC-LDPC Codes by Cyclic Group Decomposition," 2019 IEEE East-West Design Test Symposium (EWDTS), 2019, pp. 1-5

[68] Usatyuk V. Short-code performance under soft decision decoder https://github.com/Lcrypto/short-block-linear-code-soft-decision

[69] Y. Polyanskiy, H. V. Poor and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," in IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2307-2359, May 2010

[70] Numerical code for finite-blocklength research: "SPECTRE: Short packet communication toolbox," GitHub repository, Dec 2014. https://github.com/yp-mit/spectre.git

Fast RLS algorithms in Combined Adaptive Array and Fractionally-Spaced Feed-Forward/Feed-Backward Equalizer

Victor Djigan
Department of Integrated Circuits Design Methodology
Institute for Design Problems in Microelectronics of RAS
Moscow, Russia
djigan@ippm.ru

***Abstract*—This paper considers the usage of the fast Recursive Least Squares (RLS) adaptive filtering algorithms for the weights calculation in the combined adaptive unit, which consists of an antenna array and a channel equalizer. The weights of the fractionally-spaced Feed-Forward (FF) filter of the equalizer are split and used simultaneously as the antenna array weights. The array output signal is combined with that of the Feed-Backward (FB) filter of the equalizer. Because the FF filter operates at the up-sampled rate and the FB filter of the equalizer operates at the symbol rate, the multichannel fast RLS algorithms cannot be used in the considered adaptive unit directly. It is proposed to use a polyphase representation of the FF filter to solve this problem. In this case, the architecture of the adaptive unit is presented as a multichannel adaptive filter with unequal number of weights in channels. The paper considers the architecture of the adaptive unit and the details of its FF part. The mathematical details of the multichannel fast RLS algorithm (Fast Kalman, FK), used for the adaptive unit weights calculation, are presented. The simulation of the array/equalizer demonstrates the ability of the proposed unit to steer the main lobe of the antenna array radiation pattern towards the desired signal source without a priory search of the source angular location and, at the same time, to remove the signals of the external interferences and the intersymbol interference in the array output signal. The simulation has been carried out, using a linear array with 16 omnidirectional antennas. It operates in the condition of –30 dB Signal-to-Interference Ratio and 10…30 dB Signal-to-Noise Ratio. The array receives the Phase Shift Keying (8-PSK) desired signal, passed through a two-rays communication channel with about –65 dB gaps in the channel amplitude-frequency response.**

Keywords— Recursive Least Squares (RLS) adaptive filtering algorithms, Fast Kalman (FK) algorithm, adaptive antenna array, equalizer, interference, multipath

I. INTRODUCTION

The adaptive signal processing [1] is widely used in the design of the modern communication systems. One of the adaptive signal processing usage is to remove the signals of the external interference sources from the Antenna Array (AA) output signal. The arrays, which deal with this task, are called adaptive [2]. However, an Adaptive AA (AAA) operates efficiently if there is no multipath propagation of the desired signal, received by the array. In the multipath conditions, the AAA output signal contains an additional noise, caused by the intersymbol interference of the desired signal. Because of the noise, the performance of the AAA degrades. The intersymbol interference is usually removed by the additional Feed-Forward (FF), Feed-Backward (FB) or both FF and FB equalizes [3] in cascade with AAA.

The independent operation of the AAA and the adaptive equalizer usually provides a limited improvement of the communication system performance [4]. It is because of the intersymbol interference, that does not allow to achieve a perfect suppression of the signals of the external interference sources. As a result, the un-suppressed signals of the interferences at the AAA output produce an additional noise at the adaptive equalizer input that restricts the equalizer performance.

An advanced solution of the considered problem is presented in [5]. The key feature of the solution is the consideration of the AAA and the adaptive FF/FB equalizer as an indivisible adaptive unit, which weights are calculated by a common adaptive filtering algorithm. The unit [5] operates at a symbol rate. In this case, any version of the multichannel adaptive filtering algorithms [1] can be used for the AA/equalizer weight calculation: quadratic or linear (fast) complexity RLS algorithms or the algorithms based on the gradient search strategies.

Usually, most of the equalizers operate at a symbol rate, because the Symbol-Spaced (SS) equalizers require the smallest resources for their implementations. However, the SS sampling of the equalizer input signal does not satisfy the Nyquist criteria because of the aliasing phenomenon, that adds an additional noise to the in-band signal. However, this noise is tolerable if the received data symbols and analog-to-digital converter samples are synchronized.

This synchronization is not required in the Fractionally-Spaced (FS) equalizers [6]. For them, the input signal is sampled at a few times (usually it is an integer value) higher rate than the symbol rate. Such equalizers demonstrate a better performance comparing to the SS ones. The increased implementation resources (hardware, software or both) are the costs of this. It is because the number of the weights of the FF filter of the equalizer is increased linearly with the input signal sampling rate growing.

The using of the AA, FS FF and SS FB equalizers in the adaptive unit, similar to [5], is presented in [7]. Such FS units

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

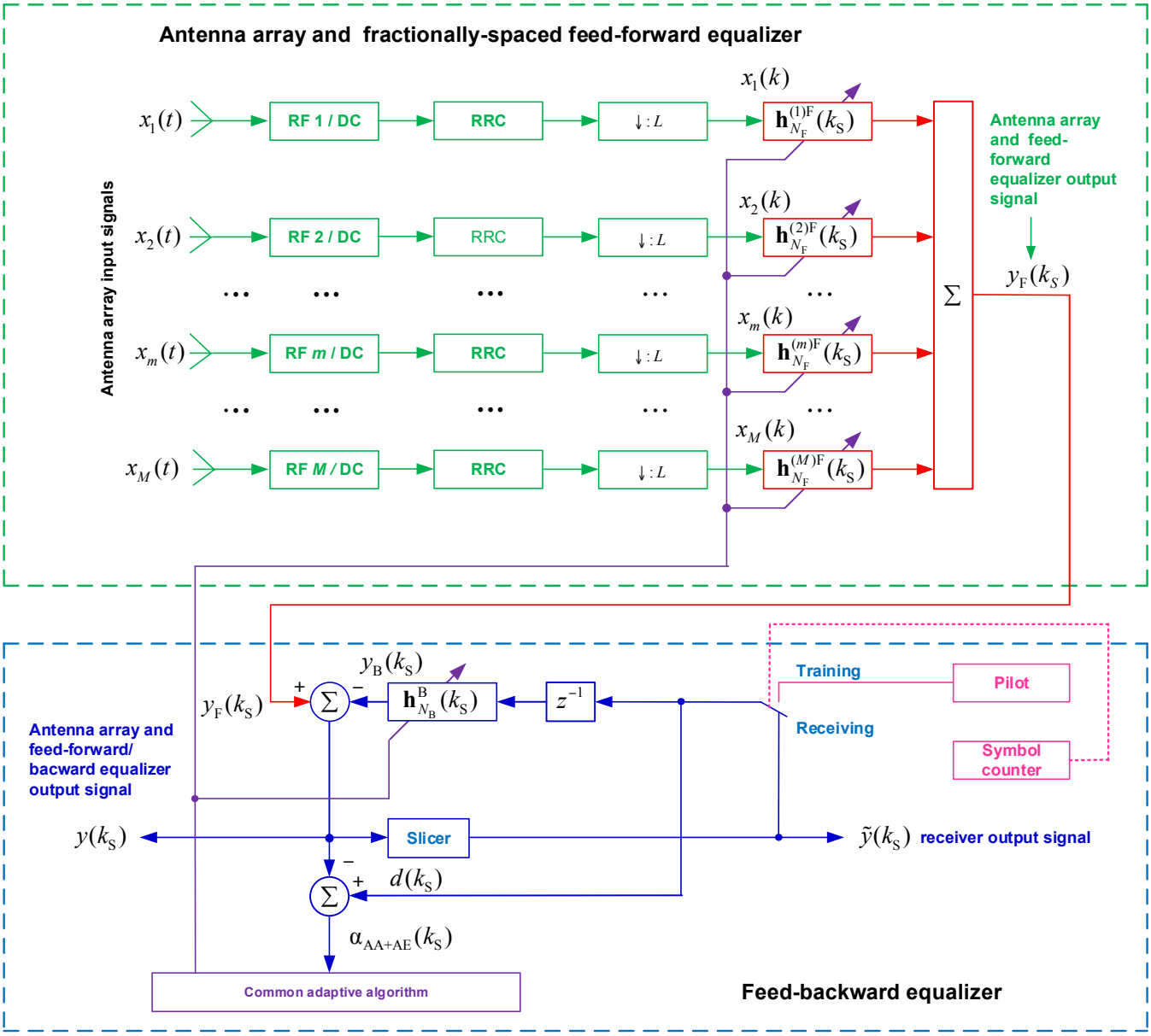


Fig. 1. Architecture of AA and FS FF/FB equalizer unit

may use almost the same multichannel adaptive filtering RLS algorithms with $O(N^2)$ complexity as those of the SS units. Here, N is the total number of the weights of the adaptive unit. However, because the FS architecture has an increased arithmetic complexity, the usage of the efficient RLS adaptive filtering algorithms with quadratic complexity $O(N^2)$ is limited in the case, when N is a large value.

The complexity may be decreased, if for the weight calculation in the unit (see Fig. 1) the mathematically equivalent multichannel fast RLS algorithms with the linear $O(N)$ complexity are used instead of the quadratic complexity $O(N^2)$ RLS algorithms [1]. However, the fast RLS algorithms cannot be used in the unit [7] directly.

The operation of these algorithms is based on the linear prediction of the adaptive filter input signal, which samples have to follow with the same rate as that of the weight update. To overcome the similar problem in equalizers, in [8] it has been proposed to use the polyphase representation of the FF filter of the FS equalizers.

The goal of this paper is to show how to use a polyphase representation of the FS AA/FF equalizer channel filters in the AAA/FF/FB equalizer and how to use the fast versions of the multichannel RLS algorithms for the weights calculation of the unit.

The new architecture of the AAA and FS FF/FB equalizer and an example of the multichannel fast RLS algorithm, called Fast Kalman (FK), are presented in the following section.

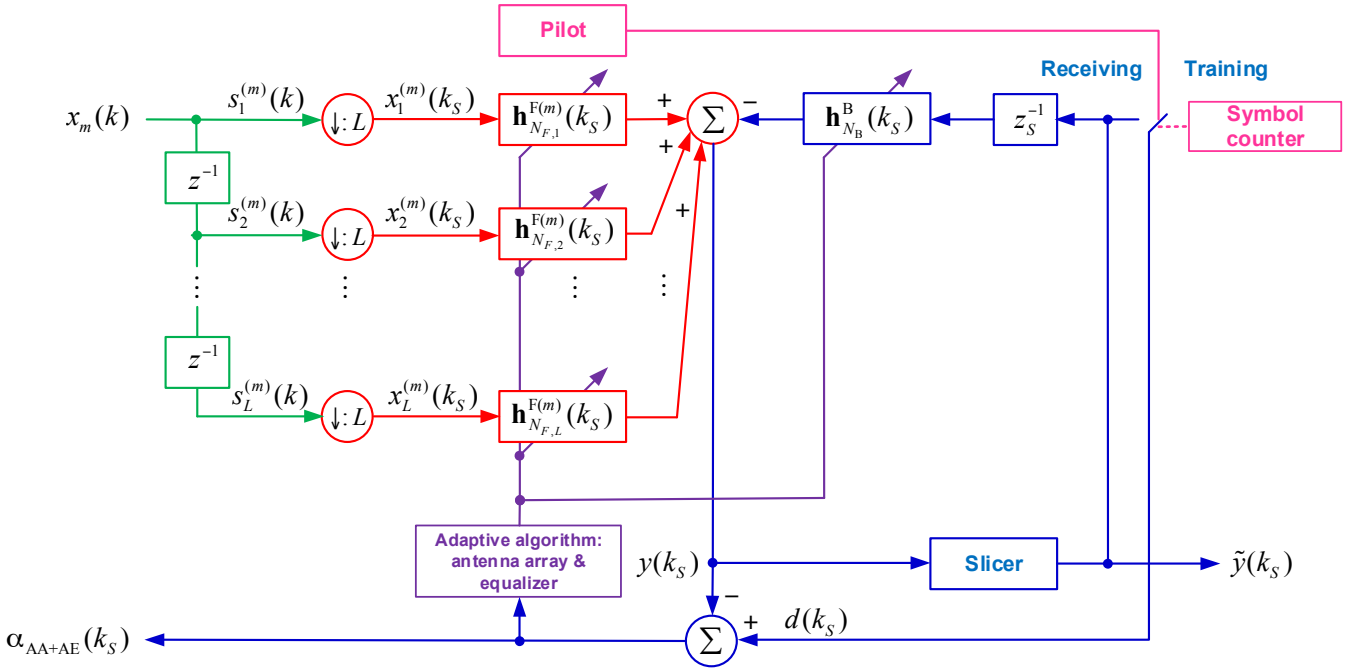


Fig. 2. Architecture of AA and FS FF/FB equalizer unit: details for one channel of the array

II. Proposed Adaptive Architecture and Algorithm for Weights Calculation

The proposed architecture of the unit (see Fig. 1) assumes the usage of the AA with the Digital Beamforming (DBF) [9]. The AA with M antennas/channels may have an arbitrary geometric configuration. Each channel has an antenna, which receives all incoming signals. It also has a Radio Frequency (RF) part, which makes the frequency selection and amplification of the signals. Besides, it has a Down-Converter (DC), which converts the signals to Base-Band (BB), and it has a Root-Riced Cosine (RRC) filter and a decimator, which are the standard blocks of any communication modem [10]. In Fig. 1, the decimator is denoted as $\downarrow : L$. Here, L is the decimation parameter of the input signal. In the FS operation, the relationship $L < I$ is established. Here, I is the interpolation parameter of the transmitter's modem.

After the frequency selection, amplification, down conversion and decimation, the AA input signals $x_m(t)$ appear in the discrete form $x_m(k)$ for the BB processing. Here, t is the continuous time and k is the discrete-time sample number.

The weights of the FS FF filter of the equalizer are split and used simultaneously as the weights of the AA. In Fig. 1, the vector $\mathbf{h}_{N_\mathrm{F}}^{(m)\mathrm{F}}(k_\mathrm{S})$ is the vector of the weights in the m-th channel of the AA, k_S is the discrete-time data symbol number, which denotes the weights update at a symbol rate.

In this paper, the lowercase characters denote the scalar variables and the elements of the vectors and matrices. The vectors and matrices are denoted by the bold lowercase and uppercase characters. The superscript T denotes the transposition of a vector or a matrix and the superscript H denotes the Hermitian transpose, i.e. transposition of a vector or a matrix and complex conjugation of its elements, denoted as $*$. The subscript indicates the number of the elements N in a vector or indicates $N \times N$ elements in a square matrix.

The FB filter output signal $y_\mathrm{B}(k_\mathrm{S})$ is combined with the AA/FF equalizer output signal $y_\mathrm{F}(k_S)$, producing at the symbol rate the signal

$$y(k_S) = y_\mathrm{F}(k_S) + y_\mathrm{B}(k_S) \tag{1}$$

of the combined adaptive unit (see Fig. 1). Here,

$$y_\mathrm{F}(k_\mathrm{S}) = \sum_{m=1}^{M} \mathbf{h}_{N_\mathrm{F}}^{\mathrm{F}(m)\mathrm{H}}(k_\mathrm{S}-1)\mathbf{x}_{N_\mathrm{F}}^{(m)}(k_\mathrm{S}) \tag{2}$$

and

$$y_\mathrm{B}(k_\mathrm{S}) = \mathbf{h}_{N_\mathrm{B}}^{\mathrm{BH}}(k_\mathrm{S}-1)\mathbf{x}_{N_\mathrm{B}}(k_\mathrm{S}), \tag{3}$$

where

$$\mathbf{x}_{N_\mathrm{F}}^{(m)}(k_S) = \left[x_m(k), x_m(k-1), \ldots, x_m(k-n_\mathrm{F}), \ldots, x_m(k-N_\mathrm{F}+1)\right]^\mathrm{T}\Big|_{\mathrm{mod}_L(k)=0}, \tag{4}$$

and

$$\mathbf{x}_{N_\mathrm{B}}(k_\mathrm{S}) = \left[d(k_\mathrm{S}-1), d(k_\mathrm{S}-2), \ldots, d(k_\mathrm{S}-n_\mathrm{B}), \ldots, d(k_\mathrm{S}-N_\mathrm{B})\right]^\mathrm{T} \tag{5}$$

are the vectors of the signals in the FS FF and SS FB parts of the equalizer, $d(k_S)$ is the desired signal (training or from Slicer output).

As it has already been mentioned, the computation of the weight vector

$$\mathbf{h}_N(k_S) = \left[\mathbf{h}_{N_F}^{F(1)T}(k_S), \mathbf{h}_{N_F}^{F(2)T}(k_S), \ldots, \mathbf{h}_{N_F}^{F(m)T}(k_S), \ldots, \mathbf{h}_{N_F}^{F(M)T}(k_S), \mathbf{h}_{N_B}^{BT}(k_S)\right]^T \quad (6)$$

of the adaptive unit (see Fig. 1), cannot be executed by the fast RLS algorithms. To use the algorithms, it is necessary to ensure the input signal shift and the weight update at the same symbol rate in the FF and FB filters. For that, the FS FF part of the unit (see Fig. 1) has to be modified in each m-th channel as it is shown in Fig. 2. In this case, the weights of each $\mathbf{h}_{N_F}^{F(m)}(k_S)$ vector are split among the L additional channels. If N_F (the same for each m-th channel) and L are odd numbers, then the vectors $\mathbf{h}_{N_F,l}^{F}(k_S)$ have the same $N_{F,l} = N_F/L$ number of weights. Otherwise, the values of $N_{F,l}$ are different.

As usually the value $N_B \neq N_{F,l}$, then it is required to use the fast versions of the multichannel RLS algorithms for the adaptive filters with the unequal number of the weights in $\hat{M} = M \cdot L + 1$ channels for the weight calculation in the unit (see Fig. 1). The multichannel algorithms [1] can be used for this purpose. An example of the computational procedure of such algorithm is presented below.

Multichannel FK RLS algorithm of combined AA/FF/FB equalizer

Initialization : $E^{f(\hat{m})}(0) = \delta^2$;

$\mathbf{h}_N^{f(\hat{m})}(0) = \mathbf{0}_N$; $\mathbf{h}_N^{b(\hat{m})}(0) = \mathbf{0}_N$; **create :** $\mathbf{S}_{N+1}^{(\hat{m})}\mathbf{T}_{N+1}^{(\hat{m})T}$;

$\hat{m} = 1, \ldots, \hat{M}$; $\hat{M} = M \cdot L + 1$; $\mathbf{g}_N^{(\hat{M})}(1) = \mathbf{0}_N$,

$\mathbf{x}_N^{(0)}(0) = \mathbf{0}_N$; $\mathbf{h}_N(0) = \mathbf{0}_N$;

$\mathbf{s}_L^{(m)}(0) = \mathbf{0}_L$; $m = 1, \ldots, M$; $l = 1, \ldots, L$; $k_s = 0$;

For $k = 1, 2, \ldots, K$

For $m = 1, 2, \ldots, M$

$\mathbf{s}_L^{(m)}(k)\big|_{2:L} = \mathbf{s}_L^{(m)}(k)\big|_{1:L-1}$, $\mathbf{s}_L^{(m)}(k)\big|_1 = x_m(k)$

End for m

$l = \mathrm{mod}_L(k)$

if $l = 0$

$k_s = k_s + 1$

For $m = 1, 2, \ldots, M$

For $l = 1, 2, \ldots, L$

$\hat{m} = m \cdot l$

$x_{\hat{m}}(k_S) = x_l^{(m)}(k_S) = s_l^{(m)}(k)$ **(decimation)**

$\mathbf{x}_{N_{F,\hat{m}}}^{(\hat{m})}(k_s)\big|_{2:N_{F,\hat{m}}} = \mathbf{x}_{N_{F,\hat{m}}}^{(\hat{m})}(k_s)\big|_{1:N_{F,\hat{m}}-1}$

$\mathbf{x}_{N_{F,\hat{m}}}^{(\hat{m})}(k)\big|_1 = x_{\hat{m}}(k_S)$

End for l

End for m

$\mathbf{x}_{N_B}(k_S)\big|_{2:N_B} = \mathbf{x}_{N_B}(k_S)\big|_{1:N_B-1}$, $\mathbf{x}_{N_B}(k_S)\big|_1 = d(k_S - 1)$

$\mathbf{x}_N^{(0)}(k_S) = \left[\mathbf{x}_{N_{F,1}}^{(1)T}(k_S), \mathbf{x}_{N_{F,2}}^{(2)T}(k_S), \ldots, \mathbf{x}_{N_{F,\hat{m}}}^{(\hat{m})T}(k_S), \mathbf{x}_{N_{F,\hat{m}}}^{(\hat{m}+1)T}(k_S), \ldots, \mathbf{x}_{N_{F,\hat{M}-1}}^{(\hat{M}-1)T}(k_S), \mathbf{x}_{N_B}^{T}(k_S)\right]^T = \left[\mathbf{x}_{N_1}^{(1)T}(k_S), \mathbf{x}_{N_2}^{(2)T}(k_S), \ldots, \mathbf{x}_{N_{\hat{m}}}^{(\hat{m})T}(k_S), \mathbf{x}_{N_{\hat{m}+1}}^{(\hat{m}+1)T}(k_S), \ldots, \mathbf{x}_{N_{\hat{M}-1}}^{(\hat{M}-1)T}(k_S), \mathbf{x}_{N_{\hat{M}}}^{(\hat{M})T}(k_S)\right]^T$

$\vdots$

$\mathbf{x}_N^{(\hat{m})}(k_S) = \left[\mathbf{x}_{N_1}^{(1)T}(k_S - 1), \mathbf{x}_{N_2}^{(2)T}(k_S - 1), \ldots, \mathbf{x}_{N_{\hat{m}}}^{(\hat{m})T}(k_S - 1), \mathbf{x}_{N_{\hat{m}+1}}^{(\hat{m}+1)T}(k_S), \ldots, \mathbf{x}_{N_{\hat{M}-1}}^{(\hat{M}-1)T}(k_S), \mathbf{x}_{N_{\hat{M}}}^{(\hat{M})T}(k_S)\right]^T$

$\vdots$

$\mathbf{x}_N^{(\hat{M})}(k_S) = \left[\mathbf{x}_{N_1}^{(1)T}(k_S - 1), \mathbf{x}_{N_2}^{(2)T}(k_S - 1), \ldots, \mathbf{x}_{N_{\hat{m}}}^{(\hat{m})T}(k_S - 1), \mathbf{x}_{N_{\hat{m}+1}}^{(\hat{m}+1)T}(k_S - 1), \ldots, \mathbf{x}_{N_{\hat{M}-1}}^{(\hat{M}-1)T}(k_S - 1), \mathbf{x}_{N_{\hat{M}}}^{(\hat{M})T}(k_S - 1)\right]^T$

For $\hat{m} = \hat{M}, \hat{M} - 1, \ldots, 1$

$\alpha^{f(\hat{m})}(k_S) = x_{\hat{m}}(k_S) - \mathbf{h}_N^{f(\hat{m})H}(k_S - 1)\mathbf{x}_N^{(\hat{m})}(k_S)$

$\alpha^{b(\hat{m})}(k_S) = x_{\hat{m}}(k_S - N_{\hat{m}}) - \mathbf{h}_N^{b(\hat{m})H}(k_S - 1)\mathbf{x}_N^{(\hat{m}-1)}(k_S)$

$\mathbf{h}_N^{f(\hat{m})}(k_S) = \mathbf{h}_N^{f(\hat{m})}(k_S - 1) + \mathbf{g}_N^{(\hat{m})}(k_S)\alpha^{f(\hat{m})*}(k)$

$e^{f(\hat{m})}(k_S) = x_{\hat{m}}(k_S) - \mathbf{h}_N^{f(\hat{m})H}(k_S - 1)\mathbf{x}_N^{(\hat{m})}(k_S)$

$E^{f(\hat{m})}(k_S) = \lambda E^{f(\hat{m})}(k_S - 1) + e^{f(\hat{m})}(k_S)\alpha^{f(\hat{m})*}(k_S)$

$\overline{\mathbf{g}}_{N+1}^{(\hat{m})}(k_S) = \begin{bmatrix} 0 \\ \mathbf{g}_N^{(\hat{m})}(k_S) \end{bmatrix} + \begin{bmatrix} 1 \\ -\mathbf{h}_N^{f(\hat{m})}(k_S) \end{bmatrix} \frac{e^{f(\hat{m})}(k_S)}{E^{f(\hat{m})}(k_S)}$

$\tilde{\mathbf{g}}_{N+1}^{(\hat{m})}(k_S) = \mathbf{S}_{N+1}^{(\hat{m})}\mathbf{T}_{N+1}^{(\hat{m})T}\overline{\mathbf{g}}_{N+1}^{(\hat{m})}(k_S) = \begin{bmatrix} \tilde{\mathbf{q}}_N^{(\hat{m})}(k_S) \\ \tilde{q}^{(\hat{m})}(k_S) \end{bmatrix}$

$\mathbf{g}_N^{(\hat{m}-1)}(k_S) = \frac{\tilde{\mathbf{q}}_N^{(\hat{m})}(k_S) + \mathbf{h}_N^{b(\hat{m})}(k_S - 1)\tilde{q}^{(\hat{m})}(k_S)}{1 - \alpha^{b(\hat{m})*}(k_S)\tilde{q}^{(\hat{m})}(k_S)}$

$\mathbf{h}_N^{b(\hat{m})}(k_S) = \mathbf{h}_N^{b(\hat{m})}(k_S - 1) + \mathbf{g}_N^{(\hat{m}-1)}(k_S)\alpha^{b(\hat{m})*}(k_S)$

End for $\hat{m}$

$$y(k_{\mathrm{S}}) = \mathbf{h}_N^{\mathrm{H}}(k_{\mathrm{S}}-1)\mathbf{x}_N^{(0)}(k_{\mathrm{S}})$$

$$\alpha_{\mathrm{AA+AE}}(k_{\mathrm{S}}) = d(k_{\mathrm{S}}) - y(k_{\mathrm{S}})$$

$$\mathbf{h}_N(k_{\mathrm{S}}) = \mathbf{h}_N(k_{\mathrm{S}}-1) + \mathbf{g}_N^{(0)}(k_{\mathrm{S}})\alpha^*_{\mathrm{AA+AE}}(k_{\mathrm{S}}) =$$

$$= \left[\mathbf{h}_{N_{\mathrm{F},1}}^{\mathrm{F}(1)\mathrm{T}}(k_{\mathrm{S}}), \mathbf{h}_{N_{\mathrm{F},2}}^{\mathrm{F}(2)\mathrm{T}}(k_{\mathrm{S}}), \ldots, \mathbf{h}_{N_{\mathrm{F},\hat{m}}}^{\mathrm{F}(\hat{m})\mathrm{T}}(k_{\mathrm{S}}), \mathbf{h}_{N_{\mathrm{F},\hat{m}+1}}^{\mathrm{F}(\hat{m}+1)\mathrm{T}}(k_{\mathrm{S}}),\right.$$

$$\left.\ldots, \mathbf{h}_{N_{\mathrm{F},\hat{M}}}^{\mathrm{F}(\hat{M}-1)\mathrm{T}}(k_{\mathrm{S}}), \mathbf{h}_{N_{\mathrm{B}}}^{\mathrm{BT}}(k_{\mathrm{S}})\right]^{\mathrm{T}}$$

$$\mathbf{g}_N^{(\hat{M})}(k_{\mathrm{S}}+1) = \mathbf{g}_N^{(0)}(k_{\mathrm{S}})$$

End for if

End for k

Here, δ^2 is the parameter, which sets the initial values of the energies of the linear prediction $E_N^{f(\hat{m})}(0)$, λ is the forgetting parameter and $\mathbf{S}_{N+1}^{(\hat{m})}$, $\mathbf{T}_{N+1}^{(\hat{m})\mathrm{T}}$ are the square permutation matrices [4]. The notation like $\bullet\big|_{n_1:n_2}$ denotes the numbers of the used elements in a vector. Besides, m and l values in Fig. 1 and Fig. 2 correspond to the values $\hat{m} = m \cdot l$ in the notations in the above algorithm. The considered adaptive filtering algorithm is easily modified to other sorts of the fast RLS algorithms: Fast Transversal Filter (FTF), Fast a Posteriori Sequential Technique (FAEST) and stabilized FAEST in the part of the computations of a posteriori Kalman gain vector $\mathbf{g}_N(k_{\mathrm{S}})$ via a priori Kalman gain vector $\mathbf{t}_N(k_{\mathrm{S}})$ and a posteriori and a priori error ratio $\varphi_N(k_S)$ [1].

III. Simulation results

The efficiency of the unit (see Fig. 1 and Fig. 2) has been demonstrated by means of simulation. A linear AA with $M = 16$ antennas and half wavelength antennas spacing has been used. The number of weights has been selected as $N_{\mathrm{F}}^{(m)} = 11$ and $N_{\mathrm{B}} = 5$. The weights have been calculated by means of the considered multichannel FK adaptive filtering algorithm. The oversampling of the input signals has been selected as $I/L = 2$. The 8-PSK (Phase Shift Keying) modulated signal has been passed through a two-ray communication channel with about –65 dB gaps in the channel amplitude-frequency response and has been received by the AA. Up to four interferences with –30 dB Signal-to-Interference Ratio (SIR) each have been simulated. The AA channel Signal-to-Noise Ratio (SNR) has been specified in the range of 10 …30 dB. The main beam of the AA radiation pattern has been initially steered towards $\theta_{\mathrm{Init}} = 33.8^\circ$ direction, where one of the interference sources was also located. All interference sources were located at $\theta_{\mathrm{Interf}} = \pm 10.3^\circ$ and $\theta_{\mathrm{Interf}} = \pm 33.8^\circ$ directions.

Fig. 3a) demonstrates the ability of the considered adaptive unit to provide the reorientation of the radiation pattern main lobe from the θ_{Init} to the θ_{Data} direction and to suppress the external and intersymbol interferences simultaneously. It happens because no distortion of the main lobe and deep gaps towards the interference sources (~100 dB) are observed in the steady-state radiation pattern. Fig. 3b) demonstrates the learning curves in terms of the radiation pattern values towards all the receiving signals. They show that the transient response in the considered case takes less than 500 symbols, which duration was specified as the duration of the training sequence. Fig. 3c) and 3d) demonstrate the quality of the communication channel equalization. The ripple of the equalized amplitude-frequency response (channel and equalizer in cascade) is less than 0.05 dB for SNR=30 dB. The steady-state Normalized Mean Square Error at unit (see Fig. 1) output is estimated as

$$\mathrm{NMSE} = -\mathrm{SNR}_{\mathrm{dB}} - 10\lg(M) = -30 - 12 = -42 \ \mathrm{dB}, \quad (7)$$

where the first term corresponds to the AA input SNR and the second one is the SNR improvement by means of the AA due to the coherent combination of the data signals and the non-coherent combination of the channel noise signals at the AA output. The actual value of the NMSE (see Fig. 3e) is about 0.9 dB less, as due to the adaptation the relationship $\left\|\mathbf{h}_{N_F}^{F(m)}(k_{\mathrm{S}})\right\|_2 = 1$ is not valid and the coherent and non-coherent combinations of the desired signal and noise are not completely ensured. Fig. 3f) demonstrates the quality of the architecture (see Fig. 1 and Fig. 2) in the terms of data constellation. It is easy to see that the channel output signal constellation is concentrated close to the data symbols after the equalization.

IV. Conclusion

Thus, the paper presents a combined adaptive AA/ FS FF/FB equalizer for signal receiving in multipath and interference conditions. It can be applied in modern wireless communication systems equipment, which use the AA with DBF as a directional antenna.

References

[1] V.I. Djigan, Adaptive filtering: theory and algorithms. Moscow: Technosphera Publisher, 2013, 528 p. (in Russian)

[2] J.E. Hudson, Adaptive array principles.The Institution of Engineering and Technology, 2007, 253 p.

[3] S. Qureshi**,** "Adaptive equalization," IEEE Communications Magazine, vol. 20, pp. 9–16, March 1982.

[4] J.-Y. Lee and H. Samueli, "Adaptive antenna arrays and techniques for high bit-rate QAM receivers," IEEE Journal on Selected Areas in Communication, vol. 17, pp**.** 677–688, April 1999**.**

[5] V.I. Djigan, "Adaptive antenna array for operation in interference and multipath conditions," Digital Signal Processing, pp. 20–27, No. 4, 2019. (in Russian)

[6] J.R. Treichler, I. Fijalkow and C.R. Johnson, "Fractionally spaced equalizers", IEEE Signal Processing Magazine, vol. 12, pp. 65–81, May 1996.

[7] V.I. Djigan, "Adaptive antenna array, shared with adaptive equalizer", Proceedings of the International Conference on Antennas Theory and Technique, Kharkiv, Ukraine, June 22 – 27, 2020 (will be published).

[8] V.I. Djigan, " Fractionally spaced feed-backward equalizers, based on fast RLS adaptive filtering algorithms", Proceedings of the All-Russian Scientific and Technical Conference on Problems of Prospective Micro and Nanoelectronic Systems Development, Moscow, Russia, Part 2, pp. 126–131, 2020. (in Russian)

[9] J. Litva and T.K.-Y. Lo, Digital beamforming in wireless communications. Artech House., 1996, 301 p.

[10] J.G. Proakis and M. Salehi, Digital communications, 5-th ed. McGraw Hill, 2007, 1170 p**.**

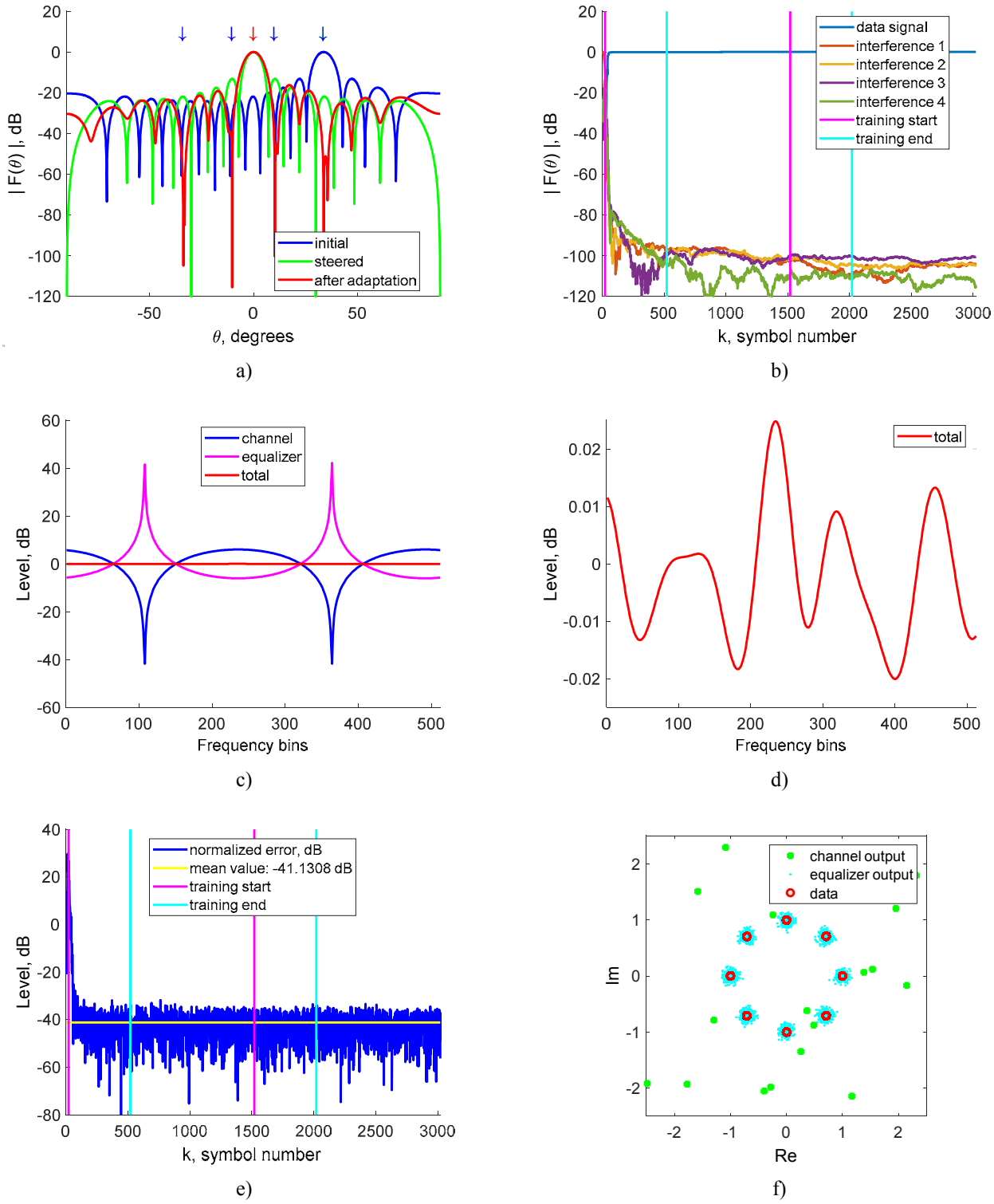


Fig. 3. Simulation results: a) is the steady-state radiation pattern, b) is transient response (radiation pattern values towards the desired signal and interference sources), c) and d) are the steady-state amplitude-frequency responses, e) is the transient response in terms of output errors, f) signal constellations (SNR=10 dB)

Quantum Deterministic Computing

Wajeb Gharibi
University of Missouri-Kansas City
MO, USA
gharibiw@umkc.edu

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

Ka Lok Man
Xi'an Jiaotong-Liverpool University
China
kalok2006@gmail.com

Svetlana Chumachenko
Design Automation Department
Kharkov National University of Radio Electronics
Kharkov, Ukraine
svetachumachenko@icloud.com

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

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

***Abstract*— A novel deterministic paradigm for creating quantum computing via photon transactions with the electrons of an atom is described. This paradigm is free of quantum logic. The practically focused evolutionary path of quantum computing starting with the classical approach is shown: Memory–Address–Transaction → Electron–Address–Transaction → Electron–Address– Quantaction (EAQ). Qubit-vector models for describing functionalities are proposed that differ from the known truth tables in the compactness of description and technological ability of the algorithm implementation for synthesis and analysis of digital devices and SoC components. For example, we consider a memory-driven algorithm for the simulation of digital devices based on qubit-vector forms for describing functionalities with a significant increase in the performance of computational processes. These processes can be utilized for analysis and synthesis through parallel execution of logic operations. A set of comparative estimates of qubit models and methods for improving the efficiency of the algorithms for simulating digital devices is presented.**

Keywords— quantum computing, quantum determinism, Memory–Address–Transaction, Electron–Address–Quantaction, quantum transactions, structure of electrons, qubit vectors, matrix data structures, digital systems-on-chips.

I. A CENTURY'S JOURNEY TO MODERN QUANTUM COMPUTING

Scientists who create a quantum computer, follow the path laid down by Emile Leon Post (1897–1954) exactly 100 years ago. Post's main practical postulate is that it is impossible to create a computer without a functionally complete basis of primitive logical functions, for example, without and–not, or–not, and-or–not, 1–xor–and [1-2]. This is true, if we proceed from the truth tables of primitive elements as the basic building blocks for the synthesis of complex circuits. However, this axiom is correct until simpler operations-transactions are found, which allow describing and/or synthesizing elementary Post logic that creates functionally complete and minimal bases of logical functions, and also any other combinational structures of arbitrary complexity. Such primitives are only two functions-transactions: read and write, which allow writing-synthesizing absolutely any computational process of any complexity. Moreover, in fact, we can consider only one operation-transaction "write-read", since the binary relation embedded in them does not exist without each other.

Otherwise, in order to write somewhere, you need to read from somewhere. This means that there is a more technological and more primitive alternative to the Post's logic and classical quantum computing, where the last one is also still following the path of quantum–mechanic–driven synthesis of controlled truth tables as basic elements [3,4]. At the same time, scientists believe that it is impossible to create a quantum computer without such mathematical tables-matrices, which require a technologically complex cryogenic mechanism for cooling a silicon chip to a level close to absolute zero.

II. THE DETERMINISTIC PATH TO QUANTUM COMPUTING

An innovative proposal is to create quantum memory-driven computing [5-7] free of quantum operations of superposition and entanglement (or, not) based on the use of characteristic equation M=Q[M(X)], specifying two write-read transactions on an atomic structure of electrons forming the computational memory M, Q, X. To eliminate two technologically complex operations from quantum computing means to significantly simplify the architecture and reduce it to a memory structure on electrons for performing transactions between them using quanta or photons. A confirmation of the consistency and validity of the proposed innovative quantum architecture can be found in publications that emphasize the steady trend towards the creation of quantum computing based on an atomic memory structure with the transfer of information using photons or quanta. The proposal under consideration is associated with the transition from the usual functional scheme of mathematics associated with elementary tables, as well as physics associated with super-cold silicon crystals. Instead, an atom with an electron is considered, the spin or orbit of which can have two states, identified with binary signals {0,1}.

But this is not the main thing. The electronics market has been under the influence of transistor logic-driven technologies for an unacceptably long time, which are moving rather slowly

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

towards microminiaturization. The situation is even worse with education in the field of computer engineering, where alternative computing, such as quantum, molecular, subatomic, is practically absent. There is no doubt that two types of computing: classical and quantum, going towards each other, will meet in 5-10 years in the deterministic structure of electrons controlled by quanta-photons. The simplest solution of the new atom-quantum computer should be based on the primitive mechanism of the read-write process, in this case {0,1} in an atomic electron to an atomic electron using a separate photon influencing on the electron – writing phase 1; and a photon emitted by an electron – reading phase 1. Thus, the creation of a new atomic elementary quantum computer is reduced to structuring a group of atomic electrons with an amount from 1 to an arbitrary finite number n, on which a single operation is performed – the write-read function using absorption or emission of a photon [4]. The classical analogue of such a computer is implemented in the MAT-structure (Memory-Address-Transaction) [1], Fig. 1.

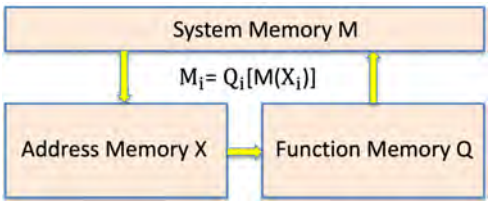


Fig. 1. Memory–Address–Transaction Computing

Here, the automatic characteristic equation of computing for organizing the computational process is free of logic: $M_i = Q_i[M(X_i)]$, where M is the memory for storing system data and the result of the computation process, Q is the memory for storing vectors-functions, access to the cells of which is provided by concatenating bits of the system memory M, which are addressed by memory cells X, creating interconnections between the components of a digital system. This logic-free structure is quite enough for the implementation and simulation of a computational process of arbitrary complexity. Atomic-quantum deterministic analogue of MAT-computing is considered below. To simplify quantum computing structurally and mathematically, excluding unnecessary logic from it, means significantly reducing the technological path (time-to-market) to obtaining a market product, where the structure and algorithm of transactions are determined by the following primitive scheme, shown in Fig. 2.

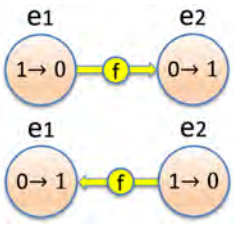


Fig. 2. Quantum transactions on the electron structure

The advantages of such computing associate with the speed of quantum (photons) transactions in the electronic memory structure, which is equal to the speed of light. It should be noted that an obstacle to conquering the market for memory-driven computing is the duration of the access cycle (read-write) to memory, which for modern silicon chips is 2 orders of magnitude longer than the signal delay for CMOS (FinFET) transistor logic. If we transfer to the structure of electrons (memory), which have addresses for writing-reading data using quanta-photons, the computational process (algorithms) on transactions between electrons becomes equal in speed to the speed of light. Hence the conclusion – the logic of the transistor loses its initial technological attractiveness in terms of performance, which becomes significantly lower than transactions in the proposed memory of the structure of electrons. The following evolution of quantum computing from the classical one arises: Memory–Address–Transaction → Electron–Address–Transaction → Electron–Address–Quantaction (EAQ). Thus, the EAQ-computing metric is formed in the following form: 1) Lack of logic that requires high energies in traditional quantum computing. 2) Light speed of photon transactions between electrons – memory elements. 3) Negligible energy consumption in the implementation of the computational process, determined by the absorption and emission of photons. 4) Determinism of all computational procedures and algorithms, determined by the addressability of electrons. 5) Addressability of data structures in the form of the order of electrons, which allows for disruptive EAQ-computing. This point represents a single problem that needs to be solved to turn quantum computing into the main engine of progress and expansion of humanity into space. 6) The synthesis of computing, thus, turns into a technology for ordering the structure of electrons at any point in the Universe. 7) Other problems of the proposed quantum EAQ-computing already have separate technological solutions focused on photonic control of atomic electrons [8-22]. 8) Quantum computing is a computational process that uses photonic transactions in the atomic structure of electrons for the parallel implementation of combinatorial algorithms in software applications [5]. A quantum computer without quantum parallel applications is just an expensive toy. Therefore, the strategy for creating quantum computing is to simultaneously develop quantum hardware and quantum software based on parallel combinatorial algorithms. The joint development of the two branches of quantum computing provides an opportunity for scientists from developing countries to actively participate in the design, simulation and verification of quantum algorithms and software applications on classical computers in order to their subsequent implementation in market-accessible quantum computers that will appear in the near future. 9) The false and practically unreal path of the material expansion of humanity into space is even theoretically the most difficult way of delivering human or a robot to planets located tens or hundreds of light years away. The correct and most realistic way is the quantum control of the teleportation process of the human, animal, plant genome for the subsequent synthesis of any organism, object or phenomenon on the planet with conditions suitable for life.

Thus, the strategy of creating an electron-address-driven quantum computer increases the relevance of the research focused on the development of unitary data structures and parallel algorithms for increasing the efficiency of solving traditional and new problems, primarily those related to combinatorics [18-22]. The proposed solutions are published in

a number of works [23-24], focused on improving the performance of algorithms and methods for designing and testing digital systems through vector-qubit data structures. Already today we can identify, three commercially viable applications for early quantum computing devices: quantum modeling, optimization, and combinatorics for the areas of health care, artificial intelligence, cyber security of data and cloud services, cryptography, transport, chemistry, and weather [25]. It is not yet known whether existing (sequential) algorithms and applications will be able to increase their performance using quantum processors that will soon be available to consumers. Therefore, today it is necessary to develop new parallel types of algorithms for quantum modeling, cryptanalysis, deep machine learning, optimization by using classical and quantum processors of companies Google, IBM, IonQ, accessible through cloud services.

In terms of energy consumption [26], the future of computing lies in the use of sunlight during the daytime phase of the day, sufficient to maintain performance. An argument for this thesis is the publication by Euisik Yoon, Sung-Yun Park in the IEEE Spectrum 2018 April magazine "Self-Powered Image Sensor Could Watch You Forever" [27]. The basic idea is that solar cells and image sensors convert light into electricity. If both components are placed on a single chip, a self-powered camera can operate by using daylight only and capable of capturing 15 images per second. A micro-system for video recording and energy harvesting, integrated with a micro-miniature processor and a wireless transceiver, makes it possible to place a small, almost invisible, camera anywhere. Integrated silicon optical chips can eliminate the problems associated with slow-response metal connections between circuit components. The optical CMOS process could break communications bottleneck [28]. Milos Popovic, professor of electrical and computer engineering at Boston University, along with colleagues from MIT, UC Berkeley, published in Nature a new way of transmitting optical signals on conventional microcircuits. The method allows one to speed up the communication between microprocessors by an order of magnitude, significantly reducing heat transfer and increasing the computing performance of laptops and smartphones. The technological idea is to add a material (a thin layer of dielectric polycrystalline silicon) on top of the existing chip components, manufactured by volumetric complementary technology. To make the material more suitable for photonics, the researchers modified the crystal structure to prevent light from leaking from their polysilicon to the substrate. The crystals obtained have all the necessary photonic components: waveguides, microcavities, vertical lattice junctions, high-speed modulators, avalanche photodetectors and transistors, manufactured using 65 nm technology. The laser source is outside the chip. Photodetectors absorb photons. The motivation for the research is that computer manufacturers are increasingly using chips and GPUs to create games and artificial intelligence that can contain hundreds of cores. At the same time, the copper wires connecting the cores are the main bottleneck holding back high performance, which, moreover, produce much heat that requires removal to the external environment. A metal conductor can transfer data from 10 to 100 gigabits per second, while optical fiber can transfer 10 to 20 terabits per second. It should also take into account that at the micro-distances between microprocessors, thermal optical losses are practically zero, so an opto-silicon system requires less energy than a copper-silicon one. The new method could lead to the creation of chips with increased processing power for hardware implementation of artificial intelligence methods for image recognition on the iPhone and in inexpensive LIDAR sensors for driver-free cars. Considering that in the next 5 years, quantum computers will appear that will require the expansion of the sequential paradigm of algorithms into parallel architectures, today it is relevant to emulate the technologies of parallel synthesis and analysis on classical computers [18, 22, 29]. Taking into account such motivation, qubit data structures and methods of their synthesis-analysis for memory-driven computing, focused on parallel execution of operations, are presented below (Fig. 3).

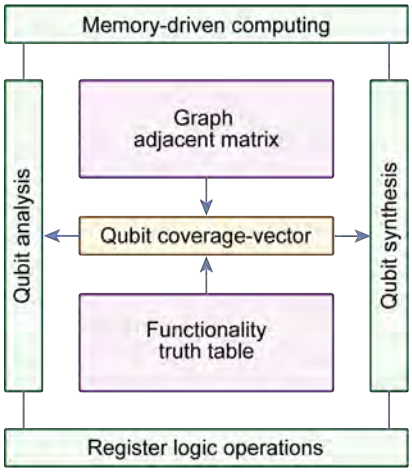


Fig. 3. The structure of qubit (quantum) computing

As a variant of using the qubit-vector model of logical elements, a simulator is further proposed that makes it possible to simulate fault-free behavior and faults in order to obtain tests for detecting faults in digital systems and components.

Data structures, transformed from a truth table to a qubit vector and a unitary coded table, and then to the execution of a parallel procedure for fault-free simulating four input sequences (1,2,4,6), are shown in Fig. 4. Here, each input sequence, represented by a binary-decimal code, is assigned a unitary code, which is characterized by the property of superposition in arbitrary combinations. This makes it possible to process in parallel an arbitrary combination of input sets for a given logical element of three variables, presented as an example.

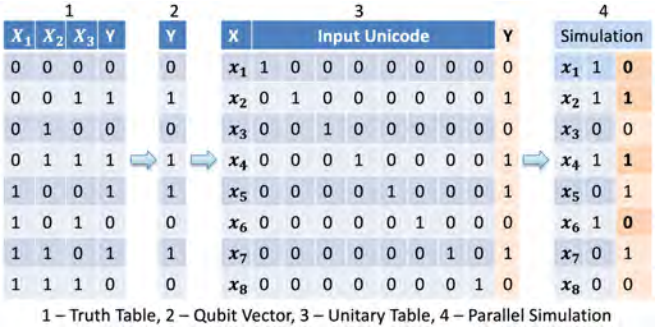


Fig. 4. Data structures and the procedure for parallel qubit simulation

In this case, simulation turns into finding the solution by means of vector-set superposition of the input word and qubit sequences that form the combinational logic of the functionality (algorithm) to achieve the goal. To get away from address identification of components in data structures as much as possible means to get high-speed performance of procedures through their parallel execution and unitary coding of set elements. Example: if there is no addressing of memory elements, which implies sequential access to cells, then non-addressable memory structures become available for writing-reading data in parallel mode. Conclusion: the addressability of data and commands is the main brake of classical computing when creating highly efficient parallel combinatorial algorithms comparable to quantum computers. Alternatively, a qubit is a set that can contain two equivalent unaddressed elements {0,1}. A qubit vector is defined as the set of qubits (binary states) that form logical functionality. A compromise solution with respect to the addressability of elements is the addressability of the set, which makes it possible to perform parallel operations with non-addressable elements. Groups of non-addressable elements are created in the following parallel-focused data structures: register and associative memory, logical combinational circuits. However, in order to perform parallel operations on the mentioned data structures, preprocessing of their readdressing is required, which takes quite a lot of time. Thus, the primitive logic of the computational process can be reduced to the formation of a relation (superposition) between a pair of operands-sets {a},{b} on the basis of a common a∪b or one of the particular metrics-operands a,b. To do this, it is necessary to transform elements from dependently addressed to positionally independent or unitary encoded. Likewise, a scientist with a greater degree of freedom can associate himself with any subset of experts, which makes him more creative compared to a specialist using an addressing. Strength or unity in diversity. In God We Trust.

The evolution of computing from sequential to quantum through parallel one is shown in Fig. 5. In the limit, quantum computing should not have any operations, except for the read–write transaction, which is already implemented in the classical sequential and parallel computing.

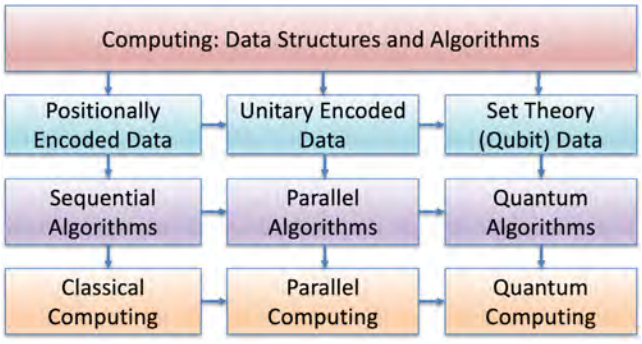


Fig. 5. Classification of computing by data structures and algorithms

The fault-free simulator has a simple and efficient interface for graphic representation of logic gates, ports of test pattern inputs and simulation outputs, as well as services for error correction and data structure storage. It is interesting that the elements here are not tied to classical logic, but they operate on qubit-vectors, which have a decimal equivalent. The structure of an example of the logic circuit is shown in Fig. 6.

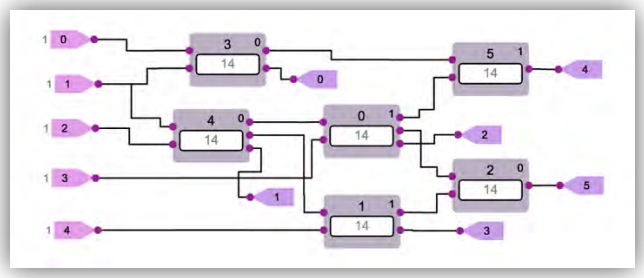


Fig. 6. Quantum simulator interface for ISCAS circuit

Inside the element there is information about the ordinal number of the primitive and the type of functionality, represented by a decimal number, which is the convolution of the qubit coverage vector of the logical element. The simulator works in a stepwise mode, when one vector is simulated, manually set on the primary inputs of the circuit. There is also an automatic exhaustive testing mode for obtaining the truth table of the entire digital device with values on all (input, internal and output) lines of the circuit. The structural organization of qubit data for simulating fault-free behavior of a digital logic circuit on test-vector is shown in Fig. 7. Here, the qubit vectors representing the elementary functions of the combinational circuit are added to the vector of fault-free simulation of a digital device. It turns out that the vertical shift of the qubit vectors relative to the horizontal simulation vector forms the states of the outputs of the digital circuit. The mathematics of this shift is defined by the characteristic simulation equation $M_i = Q_i [M(X_i)]$, which involves the horizontal simulation vector M, the qubit vectors Q, and also connection variables X, creating the cell addresses of the qubit vectors.

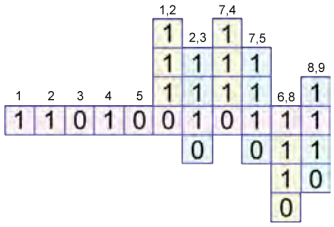


Fig. 7. Qubit data structures for a logic circuit

To obtain statistical information and verify the Quantum Modeling software application, 10 experiments were carried out on logic circuits from the ISCAS library, and also on other structures listed below: 1) Adder SP. 2) Circuit Schneider. 3) Circuit C5. 4) Circuit C17. 5) RFO Circuit. 6) MUX16 Circuit. 7) DFA Circuit. 8) Hasse processor. 9) DC4-16 Circuit. 10) Circuit C432. The comparison was made between the synthesis time of the circuit structure (Modeling Time), as well as the simulation time. The base case for comparison was the Active HDL product, Aldec Inc., where information about the circuit model was entered in the VHDL hardware description language.

The comparison statistics for Modeling and Simulation Time is shown in Fig. 8.

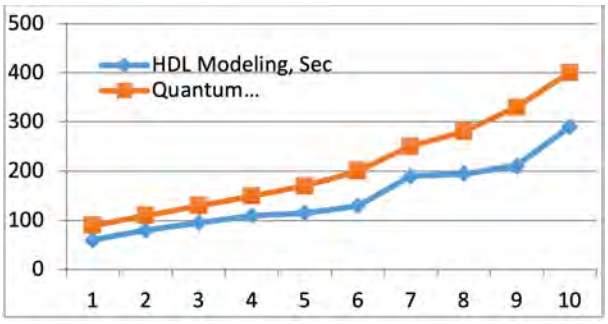


Fig. 8. Analysis of Modeling and Simulation Time

The graph shows the advantages of visual, imaginative circuit design of logic circuits of small dimensions compared to the description of circuits based on HDL-description. This is especially acceptable in the process of student education for the design and verification of digital systems and components. However, for large industrial projects, it is advisable to use hardware description languages.

III. Conclusion

A deterministic paradigm for creating quantum computing by using photonic transactions with the electrons of an atom, which excludes the use of quantum logic, is proposed. A practically focused evolutionary path of quantum computing from the classical one is shown: Memory–Address–Transaction → Electron–Address–Transaction → Electron–Address–Quantaction (EAQ). The qubit-vector models for describing functionalities are proposed, which differ from the known truth tables in the compactness of description and effectiveness in the implementation of algorithms for the synthesis and analysis of digital devices and SoC-components. Some obvious advantages of practically focused quantum computing are shown in the publications of scientists working in the fields of parallel memory-driven computing and combinatorics of information security, testing and diagnosis problems. There is no better way to advance innovative theory than the use of an open cyber-physical space where cloud services covering the interests of the general scientific community in time and space are created. Within the framework of the above mentioned scientific and technological direction, the following tools were created: 1) A graphical interface, convenient for manual design of qubit models of digital devices and components, which makes it possible to correct errors online. Alone with the high speed of modern computing, the main factor influencing the time-to-market is the time of manual input of models of digital systems and components. 2) Data structures for qubit description of digital devices and components, which are compact and highly parallel in their processing. 3) Infrastructure for the design and testing of digital devices and components with functions of storage, removal and correction of data, which makes it possible to simultaneously create digital circuits by several designers. 4) Software modules for qubit-driven simulation of digital devices and components in the modes of manual and automatic input of test sequences, which makes it possible to visually teach students the methods of synthesis and analysis. 5) Areas of future research are related to the creation of memory-driven fault models focused on qubit forms of describing functionalities with the subsequent development of a family of memory-driven parallel methods of quantum testing, simulation and diagnosis of digital systems and components.

Promising areas of research and options for their solution: 1) Any computing, usually, contains two components: memory and logic, which have various parameters for speed and cost. However, within certain limits, you can create a computer without memory or without logic. Which component, memory or logic, is technologically the most difficult in quantum implementation? Quantum computer without logic is most promising for the market in terms of hardware implementation. 2) Modern computing has a stable voltage–charge at the points of silicon space as a memory, and electrons in a subatomic space as a transactional carrier. In quantum design, electrons spin or orbit are memory, and photons are transactional carrier. To create in the future a structure of electrons that will have identifier addresses is the disruptive problem under solution. 3) The strategy of joint design of hardware and software for quantum computing is an urgent task today for technologically undeveloped countries, whose scientists can make a significant contribution to the development of parallel combinatorial algorithms and qubit-driven software applications. 4) A quantum computer efficiently and in parallel solves combinatorial problems, but in other calculations it does not give a significant gain. 5) Everything that a quantum computer does can be implemented in parallel on a classical computer with no hardware limitations. 6) In 3-5 years, a classical deterministic computer will overcome the technological barrier of 3.5 nanometers. This means that a quantum computer, moving towards the classical one from the other side (atom – electron – photon), will overcome the barrier of uncertainty and become deterministic.

References

[1] E. L. Post, Introduction to a General Theory of Elementary Propositions, Published by: The Johns Hopkins University Press, American Journal of Mathematics, 1921, Vol. 43, no. 3, pp. 163-185.

[2] P.C. Rosenbloom, Post algebras. I. Postulates and general theory, Amer. J. Math., 1942, no. 64, pp. 167-188.

[3] Z. Iftikhar, A. Anthore, A. K. Mitchell, F. D. Parmentier, U. Gennser, A. Ouerghi, A. Cavanna, C. Mora, P. Simon and F. Pierre. “Tunable quantum criticality and super-ballistic transport in a "charge" Kondo circuit,” 2018, Science 360 (6395), pp. 1315-1320.

[4] S. Sun, H. Kim, Z. Luo, G. S. Solomon, E. Waks, “A single-photon switch and transistor enabled by a solid-state quantum memory,” Science, 06 Jul 2018, vol. 361, iss. 6397, pp. 57-60.

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

[6] V.I. Hahanov, T.B. Amer, S.V. Chumachenko, E.I. Litvinova, “Qubit technology analysis and diagnosis of digital devices,” Electronic modeling, 2015, Vol. 37, no 3, pp. 17-40.

[7] V. Hahanov, W. Gharibi, E. Litvinova, M. Liubarskyi, A. Hahanova, “Quantum memory-driven computing for test synthesis,” 2017 IEEE East-West Design and Test Symposium, Novi Sad, Serbia, 2017, pp. 123-128.

[8] T. Zhong, J. M. Kindem, J. G. Bartholomew et al, “Nanophotonic rare-earth quantum memory with optically controlled retrieval”, In Science, 29 Sep 2017, vol. 357, iss. 6358, pp. 1392-1395.

[9] J. Kim, H.-Ch. Lee, K.-H. Kim et al, “Photon-triggered nanowire transistors”, In Nature Nanotechnology, no 12, 2017, pp. 963–968.

[10] G. Lovat, B. Choi, D. W. Paley et al, “Room-temperature current blockade in atomically defined single-cluster junctions”, In Nature Nanotechnology, 2017, Nov,12(11), pp. 1050-1054.

[11] Ch. Li, Zh. Wang, Y. Lu, X. Liu, L. Wang, “Conformation-based signal transfer and processing at the single-molecule level”, In Nature Nanotechnology, 2017, vol. 12, pp. 1071–1076.

[12] P. Lodahl, S. Stobbe, “Quantum photonic researchers start new company, Sparrow Quantum”, [http://www.nbi.ku.dk/english/news/news16/quantum-photonic-researchers-start-new-company-sparrow-quantum/]

[13] “IonQ Raises $20M Series B Round Led By NEA, GV To Advance Quantum Computing For Commercial Applications,“ [https://www.prnewswire.com/news-releases/ionq-raises-20m-series-b-round-led-by-nea-gv-to-advance-quantum-computing-for-commercial-applications-300494456.html]

[14] R. Hyman, “Weird quantum particles simulated in droplet of ultracold gas,” [http://www.sciencemag.org/news/2018/03/weird-quantum-particles-simulated-droplet-ultracold-gas?utm _campaign =news_daily_2018-03-06&et_rid=69734703&et_cid=1891545 By Randall Hyman. Weird quantum particles simulated in droplet of ultracold gas]

[15] N. Samkharadze, G. Zheng et al, “Strong spin-photon coupling in silicon”, In Science, 2018, vol. 359, iss. 6380, pp. 1123-1127.

[16] A. Cho, “Vibrations used to talk to quantum circuits”, In Science, 2018, vol. 359, iss. 6381, pp. 1202-1203.

[17] “The Application of Spintronics,” [http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/spintronics/]

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

[19] U. Reinsalu, J. Raik, R. Ubar and P. Ellervee, "Fast RTL Fault Simulation Using Decision Diagrams and Bitwise Set Operations," In 2011 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems, Vancouver, BC, 2011, pp. 164-170.

[20] H. Goudarzi, M.J. Dousti, A. Shafaei et al, “Design of a universal logic block for fault-tolerant realization of any logic operation in trappedion quantum circuits”, In Quantum Information Processing Journal. Springer US, vol. 13, iss. 5, 2014, pp. 1267–1299.

[21] A. J. Daley, “Quantum computing and quantum simulation with group-II atoms”, In Quantum Information Processing Journal, Dec. 10: 865, Springer US, 2011.

[22] J. B. Hill, “Leading Through Digital Disruption”, Gartner, Incorporated, 2017. [https://www.gartner.com/binaries/content/assets/events/keywords/enterprise-architecture/epaeu18/gartner_enterprise_architecture_technology_innovation_summit_digital _disruption_ebook.pdf]

[23] V. Hahanov, W. Gharibi, E. Litvinova and S. Chumachenko, "Qubit-driven Fault Simulation," In 2019 IEEE Latin American Test Symposium (LATS), Santiago, Chile, 2019, pp. 1-7.

[24] V. Hahanov, M. Liubarskyi, W. Gharibi, S. Chumachenko, E. Litvinova and I. Hahanov, "Test Synthesis for Logical X-functions," In 2018 IEEE East-West Design & Test Symposium (EWDTS), Kazan, 2018, pp. 1-9.

[25] M. Mohseni, P. Read, H. Neven, S. Boixo, V. Denchev, R. Babbush, A. Fowler, V. Smelyanskiy and J. Martinis, “Commercialize quantum technologies in five years,” [https://www.nature.com/news/commercialize-quantum-technologies-in-five-years-1.21583]

[26] A. Drozd, S. Antoshchuk, J. Drozd et. al., “Checkable FPGA Design: Energy Consumption, Throughput and Trustworthiness,” in book: Green IT Engineering: Social, Business and Industrial Applications, SSDC, vol. 171, Springer, Berlin, 2019, pp.73-94.

[27] S. K. Moore. “Self-Powered Image Sensor Could Watch You Forever” [https://spectrum.ieee.org/tech-talk/semiconductors/optoelectronics/selfpowered-image-sensor-could-watch-you-forever?utm_source=sensors&utm_campaign=sensors-04-17-18&utm_medium=email]

[28] N. Savage, “Supercharging Chips by Integrating Optical Circuits,” [https://spectrum.ieee.org/tech-talk/semiconductors/optoelectronics/optics-on-chips-could-speed-up-computing?utm_source=computingtechnology&utm_campaign= computingtechnology-05-01-18&utm_medium=email]

[29] A. Drozd, J. Drozd, S. Antoshchuk et. al., “Objects and Methods of On-Line Testing: Main Requirements and Perspectives of Development,” IEEE East-West Design & Test Symposium, Yerevan, Armenia, 2016, pp. 72-76.

Improving the Monitoring Systems Algorithmic Support for Railway Automation Equipment's Based on Dynamic Questionnaires

Dmitrii V. Efanov,
DSc, Associate Professor,
First Deputy General Director – Chief Engineer of Vega LLC,
Professor at Higher School of Transport,
Institute of Mechanical Engineering, Materials and Transport,
Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation
TrES-4b@yandex.ru

Valerii V. Khóroshev,
Ph. D. Student, Department
of "Automation, Remote Control
and Communication on Railway Transport",
Russian University of Transport (MIIT)
Moscow, Russia
Hvv91@icloud.com

***Abstract*—The paper authors consider the applied questionnaires theory. The authors proposed a questionnaires classification into static and dynamic. Static questionnaires are characterized by the constancy of their parameters (numbers of questions outcomes, questions implementation costs and events weights). Dynamic questionnaires are characterized by the questions or events changing possibility. The questionnaires classification is given. The main dynamic questionnaires for automation control systems wayside objects in railway transport are given. The dynamic questionnaires feature for railway automation devices are discussed. It is proposed to use dynamic questionnaires to improve the algorithmic support of monitoring systems for railway automation devices. Diagnostic parameters used in monitoring systems are given. Recommended diagnostic parameters sets are noted for monitoring procedures with the required completeness and diagnosis depth, as well as for solving the reliable forecasting-based problem on the obtained statistical data. It is proposed for each railway automation devices to consider operating conditions and modes and build individual dynamic questionnaires for each of the objects in the monitoring system software. An algorithm for adapting dynamic questionnaires in monitoring systems has been developed. The use of dynamic questionnaires makes it possible to improve the algorithmic support of modern monitoring systems, which, in turn, helps to increase the completeness and diagnosis depth and further forecasting reliability.**

Keywords—questionnaire theory; dynamic questionnaires; monitoring systems for railway automation devices; technical diagnosis; condition monitoring; questionnaires for wayside automation objects.

I. INTRODUCTION

The questionnaire Q is a pair of $<S, \Pi>$, where $S = \{s_1, s_2, \ldots, s_m\}$ – is the identifiable events set, and $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$ – is the questions set needed to separate events [1 – 4]. Each question in the questionnaire is characterized by three parameters: $a(\pi_i)$ – is the questions set needed to separate events; $c(\pi_i)$ – question cost; $p(\pi_i)$ – event weight. Each event in the questionnaire is characterized by a weight $p(s_j)$. The question weight is the sum of the weights identified event in the subsets of its outcomes. The question posed on any events set corresponds to the procedure for splitting it into $a(\pi_i)$ or fewer subsets. Per the values of $a(\pi_i)$ in the questionnaire, a class of homogeneous and heterogeneous questionnaires is distinguished. In homogeneous questionnaires, all questions have the same basis, and in heterogeneous – at least one reason is different from the others. In practice, homogeneous questionnaires are widely used – questionnaires with the basis of all questions $a(\pi_i) = 2$ [5 – 7]. Such questionnaires make it possible to separate the initial events set into singleton subsets with two-outcome procedures. For example, in the technical diagnosis of an object, a binary question corresponds to a check with two outcomes: "*no*" (*0*) and "*yes*" (*1*) [8]. Homogeneous questionnaires with higher base values are less common. Moreover, when posing a non-binary question on a subset of the original events set, fewer subsets of partitions can form than the value $a(\pi_i)$. Thus, a homogeneous questionnaire is converted to a heterogeneous one. In homogeneous questionnaires, as well as binary questionnaires, are widespread [9]. For example, in [10] an example, heterogeneous questionnaires in constructing optimal conditional algorithms for diagnosing technical objects of railway automation is given.

Each questionnaire is characterized by several parameters: the number of questions used to identify all events, the questions maximum number to solve the problem of event identification, the route cost and implementation cost. One of the main indicators is the implementation cost $C(Q)$. The implementation cost is defined as the average time for identifying events on the questionnaire [1]:

$$C = \sum_{j=1}^{n} p(\pi_j)\, c(\pi_j).$$

The main task of the questionnaire's theory is to build a questionnaire with a minimum value of the implementation cost. Optimization methods for questionnaires of various types are discussed in [1 – 3, 5 – 7, 11, 12].

As a rule, in all studies, questionnaires with constant parameters values $a(\pi_i)$, $c(\pi_i)$, $p(\pi_i)$, $p(s_j)$. are considered. Such questionnaires are built once for any of the applications. For example, in technical diagnostics for a specific diagnostic object according to the diagnostic model, all questions with all possible answers are received [13]. Further, the constructing process a diagnostic algorithm consists of choosing a sequence of questions necessary to solve the identifying a diagnostic event problem. We will call such questionnaires *static* since their parameters never change. In real practical problems, technical objects have variable structures, the weights of events, costs, the weight of issues, and the outcomes change with time [14]. A questionnaire arises where the parameters $a(\pi_i)$, $c(\pi_i)$, $p(\pi_i)$, $p(s_j)$ change over time. Over

978-1-7281-9899-6/20/$31.00 ©2020

time and in each specific application, the questionnaire may contain questions that allow errors in answers and uncertainties [15]. We call the questionnaires, for which at least one of the parameters may be changed, *dynamic*.

In this paper, the authors present the study's results in the field of dynamic questionnaires application in the constructing algorithms tasks for railway automation devices diagnosing. It is proposed to use dynamic questionnaires in systems of automated and automatic monitoring of railway automation devices parameters.

II. QUESTIONNAIRE CLASSIFICATION

Let us consider the classification of dynamic questionnaires by the example of their use in software for technical objects monitoring systems. Such systems include diagnostics objects, a sensor set that allows obtaining discrete and analogue diagnostic parameters, data transmission networks, as well as hardware and software for storing and processing diagnostic data [16]. Thus, the monitoring system allows for each diagnostic object to accumulate and analyze diagnostic data, which ultimately allows the identification of fault events. In this case, it is possible to identify already occurring failure events (from a finite set of them), as well as predicting their occurrence at the stages of subcritical (prefailure) states [17]. For each object of diagnosis, it is possible to accumulate a diagnostic data unique set inherent only to it, which allows even different objects of the same type to produce different results of diagnosis and prognosis.

In the railway automation field example is giving in [18]. Among all railway automation objects, the most vulnerable from the standpoint of reliability are the wayside technological equipment objects – railway track circuits and switch point machine [10, 19, 20]. At railway stations, railway track circuits and switch point machine are of the same type. Nevertheless, each of the objects of the same type is operated in different conditions. For example, there can be a different load on objects, a different influence of rolling stock, different periods of operation, etc. In modern monitoring systems for railway automation devices, these features of similar objects are not considered in any way [16, 17, 21, 22]. This circumstance does not allow a sufficiently effective monitoring procedure. The share of useful information is extremely low. The monitoring systems effectiveness can be improved by using dynamic questionnaires that are adaptable to the monitoring system software.

In Fig. 1 shows the questionnaires classification based on their features. At the first level of classification, the questionnaires are divided into static Q_S and dynamic Q_D. As noted above, static questionnaires are built once and never change, and dynamic questionnaires can change parameters over time. Dynamic questionnaires can be built for objects of the same type with general statistics on the functioning process ($Q^T{}_D$ questionnaires) and each device individually, individual dynamic questionnaires ($Q^I{}_D$ questionnaires). The latter, in turn, can be divided into two classes, depending on what initial diagnostic parameters are used in monitoring systems. These can be dynamic questionnaires, which are built for devices with diagnostic parameters already entered the system ($Q^{I0}{}_D$ type questionnaires), or alternatively questionnaires for which many diagnostic parameters are expanded ($Q^{I1}{}_D$ type questionnaires). The expansion of many diagnostic parameters in some cases is required to increase the completeness and depth of diagnosis and further forecasting.

Consider the constructing principles dynamic questionnaires for the main wayside automation devices on railway transport.

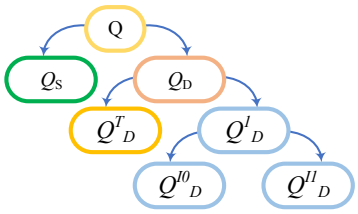


Fig. 1. Questionnaire classification.

III. DYNAMIC RAILWAY AUTOMATION QUESTIONNAIRES

A. Diagnostic models of typical railway automation devices

The main traffic control devices at stations and railroad haul are distributed state sensors of the track sections (railway track circuits), devices for automatically transferring switchblades to extreme positions (switch point machine), and devices for transmitting traffic commands to the driver (traffic lights). These objects are indicated on the schematic station's plans and stages during the design of the alarm system (Fig. 2) [23]. These objects are classified as wayside technological equipment of railway automation since they are located nearby (and railway track circuits are used to transmit current to the rails) to the railway track ("in the field"). These objects are the most vulnerable in terms of reliability. Their work substantially depends on external conditions.

Wayside technological equipment is connected to the traffic control system installed at the centralization station using a cable network deployed in cable trenches in the designated places of the station. The control system implements the interdependence between the switch points and traffic lights, allowing to organize the safe movement of trains and shunting trains [18].

Let us consider the main wayside technological railway automation devices. In this case, as an example, let consider the audio-frequency railway track circuits [17], the most common switch point machine in Russia is SP-6 (SP-6M) [10] and the lens entrance traffic light [18]. In Fig. 3 – Fig. 5 are diagnostic models of these devices.

Fig. 3 shows in the form of blocks: x_1 – track generator; x_2 – directional track filter; x_3 – track transformer of the supply end; x_4 – track jumper of the supply end; x_5 – railway line; x_6 – track jumper of the receiving end; x_7 – track transformer receiving end; x_8 – track receiver; x_9 – track relay. Some of the equipment is located at the signalling box (SB), a part is located near the railway track and is located in railway track box (TB), and also partially combined with the railway line (the signal current flows along the rails).

Switch point machine contains the following components (Fig. 4): x_1 – block contact; x_2 – electric motor; x_3 – gear; x_4 – friction cutch; x_5 – autotripper; x_6 – control rulers; x_7 – slide gear; x_8 – electric installation. All the switch point

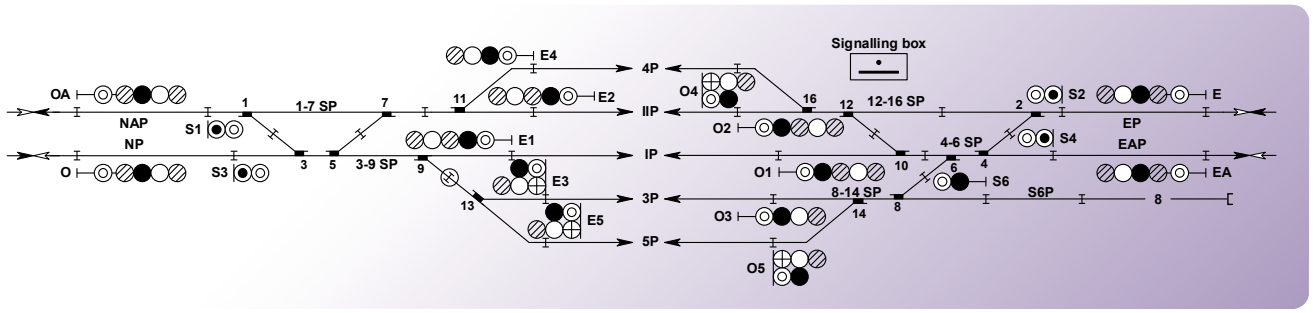


Fig. 2. Schematic plan of an arbitrary intermediate station.

machine equipment is in a castiron case, fixed with a switch mounting to the rails.

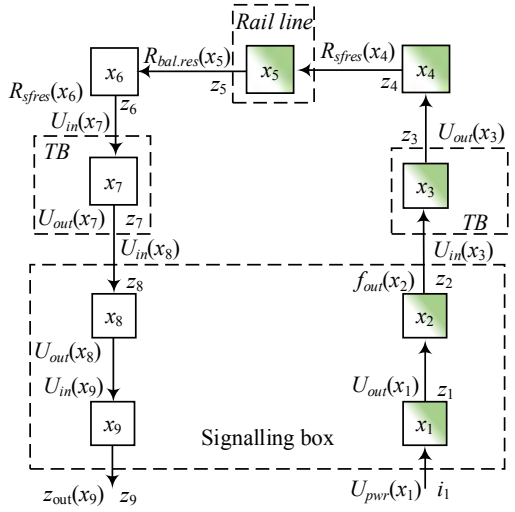


Fig. 3. Diagnostic Model of Audio-frequency railway track circuit.

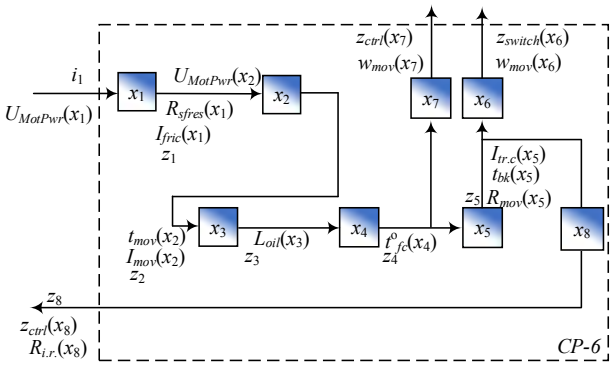


Fig. 4. Diagnostic model of switch point machine.

The traffic light includes the following objects: x_1, x_5, x_9, x_{13}, x_{17} – signal transformers (in the diagram - one for each signal indication, one for two-filament lamps - two); x_2, x_6, x_{10}, x_{14}, x_{18} – electric installation; x_3, x_7, x_{11}, x_{15}, x_{19} – lamp holder; x_4, x_8, x_{12}, x_{16}, x_{20} – traffic lights (respectively, yellow overhead light, green, red, yellow lower light and white light).

In Fig. 2 – Fig. 5, various current circuits are marked with colors for various operating modes. The presence of such operating modes makes it possible to synthesize dynamic questionnaires for these devices that operate in the monitoring system software.

In modern monitoring systems for audio-frequency track circuits, the supply voltage measurements of the track receiver and the track relay, the output voltage of the track generator, as well as the insulation resistance of the cable conductors are provided. For the switch point machine, measurements of phase currents and interphase voltages, power supply voltages of the control circuit, time of switch-blade transfer to the extreme positions, insulation resistance of the cable conductors are implemented. For a traffic light, the voltage in the working circuit and the insulation resistance of the cable conductors are measured. All these parameters in modern monitoring systems are measured not on the objects themselves distributed over the station, but on the relay cabinets of the control system located at the centralization station [16].

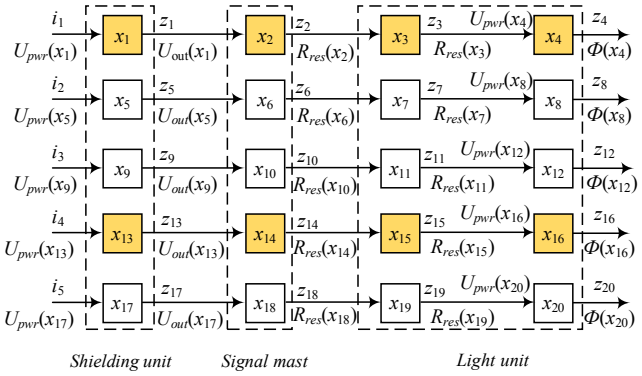


Fig. 5. Diagnostic model of the entrance traffic light.

Using the above diagnostic parameters for all objects of the same type makes it possible to build dynamic $Q^{T}{}_{D}$ type questionnaires. For such questionnaires, events weights and questions costs will use general statistics from all standard objects. Currently, in monitoring systems, the fixing deviations process from the norms is implemented upon reaching the established boundary values (normals). This is not correct since it does not consider the specific features of each diagnosis objects and leads to the fixation of many false diagnostic events. According to experts, the share of useful information from monitoring systems of railway automation in Russia is extremely low and amounts to about 5% of the total data. Using the same diagnostic parameters, $Q^{I0}{}_{D}$ type questionnaires can be synthesized. For such questionnaires, all physical properties of a diagnostic object and all specific conditions of its operation will be considered.

To increase the completeness and diagnosis depth and the subsequent forecasting reliability, it is necessary to expand the diagnostic features set. For audio-frequency railway track

circuits, measurements are also required of the current frequency at the input and output of the track generator, the voltages and currents at the inputs of the track receivers and track generators, and the resistance of the ballast of the railway prism. For switch point machines, it is necessary to measure the temperature and humidity inside the drive casing, the level and temperature of the oil in the friction clutch, the vibration effects on the gate, the mechanical parameters of the automatic switch (knife stroke), and also the geometric parameters of the moving parts of the switch (first of all, the gap between the wrench and frame rail). For a traffic light, it is necessary to measure voltages and currents in each of the switching circuits of each lamp of the lens kit, measure the angle of the mast deviation from the design axes, and also the level of vibration effects on the traffic light. Some of these parameters are currently technically difficult to measure automatically, therefore, measurements are made by maintenance personnel. To the above diagnostic parameters, weather monitoring at the station should be added using a weather station. Using the presented diagnostic parameters, $Q^{I1}{}_D$ type questionnaires can be synthesized.

It should be noted that in the future, diagnostic data should be obtained directly at each of the wayside automation objects, and not remotely from post devices. For the implementation of direct measurements, it will be necessary to install measuring controllers near the objects distributed across the stations (in track boxes, couplings, signal case and traffic light heads). Also, it is possible to combine control and monitoring functions in one design with easily removable measuring modules. Such distributed train control systems are the near future of railway transport [24].

Below are examples of questionnaires various types for the railway automation devices.

B. *Questionnaires for railway track circuits*

In Fig. 6 shows all four types of questionnaires for audio-frequency railway track circuits. The questions in it are formulated as follows:

- π_1: “Is there a slow voltage drop at the output of the track generator, a sharp decline or does not change?”;
- π_2: “Is there a ripple voltage between the lower and upper limits at the input of the track receiver?”;
- π_3: “Is the voltage at the output of the track generator increased by 5-10%?”;
- π_4: “Is the voltage at the input of the track receiver at a minimum level?”;
- π_5: “Is the input voltage to the track receiver below normal?”;
- π_6: “Is the voltage at the output of the track receiver below normal?”.

For example, describe the Q_S questionnaire for the audio-frequency railway track circuit. The root question is π_6, which separate the complete events set into two subsets: $\{s_4, s_8 - s_{10}\}$ by a single outcome and $\{s_1 - s_3, s_5 - s_7\}$ by an outcome “0”. The subset $\{s_4, s_8 - s_{10}\}$ is further divided by π_4 into two subsets: $\{s_{10}\}$ (event identification s_{10} – is a rail breakage) and $\{s_4, s_8, s_9\}$ (s_4 – are short-cut failure in the railway track circuits, s_8 – is the growth of transient resistances and breaks in cable and conductor bond, s_9 – the growth of transition resistance and breaks in track boxes and impedance bond). The subset $\{s_1 - s_3, s_5 - s_7\}$ is further divided by the question π_5 into the subsets $\{s_5 - s_7\}$ and $\{s_1 - s_3\}$. Then the first subset is broken up by the question π_3, and the second – π_2.

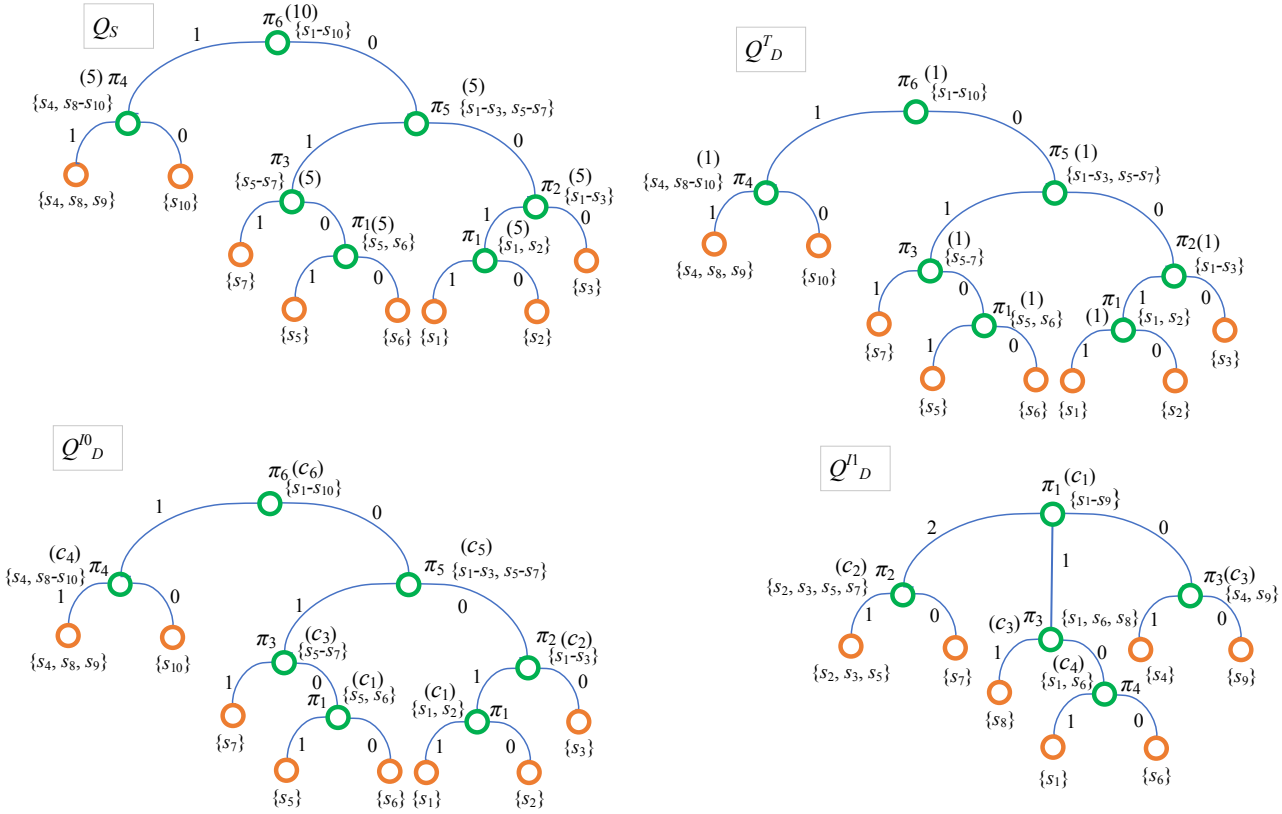


Fig. 6. Questionnaires family for audio-frequency railway track circuits diagnostic.

This makes it possible to identify the following events: s_1 – is the track generator pre-failure condition, s_2 – is the track receiver pre-failure condition, s_3 – is the pre-failure state of the track filter, s_5 – is the track generator failure, s_6 – is the track receiver failure, s_7 – is the track filter failure.

The $Q^T{}_D$ and $Q^{I0}{}_D$ questionnaires for the railway automation device under consideration differ only in the temporal characteristics of the implementation of inspections. The cardinal difference has a $Q^{I1}{}_D$ type questionnaire. It is implemented by adding new measuring points. The questions here are formulated differently, since it is possible to increase the completeness of diagnosis, and the number of identifiable events expands to nine. On the full events set, the question π_1 – is asked - "What is the type of failure?". Which automatically allows the software to separate the entire events set into three subsets, depending on. If the answer to question π_1 is the outcome "2" (a failure of the "false clear" type), then a events subset $\{s_2, s_3, s_5, s_7\}$ is formed. On this subset, for further identification of events, the question π_2, is asked, formulated as follows: "Is the voltage at the output of the track filter normal?" Question π_2 identifies events: s_2 – third-party feeding of the track receiver, s_3 – low shunt sensitivity, s_5 – is the track receiver's elements. These events are identified by a single outcome. By outcome "0", the event s_7 is identified. On the events subsets that belong to the "1" and "0" outcomes of the root question, the same question π_3 – "Is the normal voltage at the track generator output?" This question allows us to split the subset of the outcome "0" of the root question into single-element subsets corresponding to the states s_4 (low ballast resistance) and s_9 (abnormal current frequency at the track generator output, which is a precautionary state of the track generator). Question π_3 asked on a subset of the single outcome of the root question, allows us to highlight the event s_8 (disconnection between the receiver and the filter), as well as the subset consisting of events s_1 (the track generator failure) and s_6 (the track generator supply failure). The last subset is divided by question π_4, formulated as follows: "Is the generator supply voltage normal?".

From the questionnaire's analysis shown in fig. 6, it follows that some of the events in them remain not divided. This requires additional checks by the railways operating personnel. Automatically, using the monitoring systems means, a few events cannot be identified. In this case, however, the time to localize the fault is reduced.

Pay attention to the fact that the Q_S и $Q^T{}_D$ questionnaires contain numerical data on the question's costs and the weight of the events since they are based on general statistical data on the technical object's operation. Questionnaires of the $Q^{I0}{}_D$ и $Q^{I1}{}_D$ types in the examples given do not operate with any numerical values, since the prices of questions and the weight of events are determined for each object individually, taking into account its location relative to the control centralization post, physical parameters, and historical operation data.

C. Questionnaires for switch point machine

Questionnaires for switch point machine are shown in Fig. 7.

The Q_S questionnaire includes the following questions:

- π_1: "What is the ammeter reading?";
- π_2: "Is there a failure of the switch autotripper after a visual inspection of all parts of the switch, control rulers and slide gear?";
- π_3: "Is a drive failure detected after a mechanical check of the control rulers mounting loosening or the slide gear, as well as the gearbox?";
- π_4: "Is the insulation resistance in normal?";
- π_5: "Is a drive failure detected after a visual inspection of the cable contact?";
- π_6: "Is the commutation and motor separately normal?".

From the questions, Q_S is built based on expert evidence. The root question is question π_1. It separates the original events set into three subsets: $\{s_1 - s_7, s_9 - s_{15}, s_{23}\}$, $\{s_{13} - s_{22}\}$ и $\{s_8\}$. The "0" outcome of the root question corresponds to the answer "The arrow of the ammeter is stationary". To an outcome "1" – the answer "The arrow of the ammeter makes a throw, then the current value is < 2A". Outcome "2" – the answer is "First, the ammeter needle makes a throw up to 5A, and during the transfer is held at 2A." By the outcome "0" of the root question, event s_8 is identified (friction clutch misregistration has occurred). Events on the outcome of "2" of the root question are separated by question π_2. This question highlights the subsets $\{s_2 - s_6, s_9, s_{11}\}$ and $\{s_1, s_7, s_{10}, s_{12} - s_{15}, s_{23}\}$. The "1" outcome of question π_2 includes an events subset s_2 (the autotripper contact pads breakage), s_3 (the autotripper knife pads breakage), s_4 (the autotripper levers breakage), s_5 (the autotripper springs breakage), s_6 (the autotripper contact springs breakage), s_9 (the control rulers breakage), s_{11} (hit by a foreign object that impedes the movement of the slide gear). Subsequent identification of events on the resulting subset is done by manual checks. The "0" outcome of the question π_2 forms the subset $\{s_1, s_7, s_{10}, s_{12} - s_{15}, s_{23}\}$, on which the question π_3. is asked. It separates this subset into $\{s_{10}, s_{12}, s_{23}\}$ and $\{s_1, s_7, s_{13} - s_{15}\}$. According to the "1" outcome, events s_{10} (the control rulers mounting misadjustment), s_{12} (the slide gear mounting misadjustment), s_{23} (the friction clutch failure) are identified. The subset "0" of the outcome is separated by the question π_4 into the subsets $\{s_1, s_7\}$ and $\{s_{13} - s_{15}\}$. The "1" outcome of question π_4 allows identifying the events s_1 (the autotripper contact pads misadjustment) and s_7 (the autotripper contact pads covered with hoarfrost).

Back to the root question. Its outcome «1» corresponds to a subset of $\{s_{13} - s_{22}\}$. It is separated by the question π_5 into the subsets: $\{s_{16}\}$ and $\{s_{13} - s_{15}, s_{17} - s_{22}\}$. The outcome «1» of this question allows to identify the event s_{16} (the cable connection failure). As a result of «0», a subset of $\{s_{13} - s_{15}, s_{17} - s_{22}\}$. remains. It is further separated by question π_6. The question, depending on the answer, detects the following events: s_{13} (hook-up wire breakdown), s_{14} (insulation fault), s_{15} (the adjusting clip failure), s_{17} (break in the stator winding of the motor), s_{18} (break in the winding of the motor armature), s_{19} (the brush assembly motor failure), s_{20} (lowering the insulation of the electric installation), s_{21} (defect of the electric motor manifold), s_{22} (hook-up wire break-down in the electric motor).

Let us pay attention to the dynamic questionnaires for the switch point machine. Note that in Fig. 7, for each type of dynamic questionnaire, two questionnaires are given. The first (upper) corresponds to the operating mode of the arrow switch and, accordingly, the active operating mode of the

engine. The second (lower) corresponds to the control mode of operation when the electric motor is switched off.

A typical Q^T_D dynamic questionnaire is implemented as follows. The upper questionnaire differs only in the cost of the questions since the use of a monitoring system reduces the time for posing questions. An example of a change in the check implementation time is π_3. For example, in the Q_S questionnaire, the cost of the question is marked as 25 minutes, and in the Q^T_D – questionnaire as 10 minutes. The reason is that in the Q^T_D questionnaire, the travel time to the location of the drive was deleted since the check π_3 is performed after the check π_2 which is already done at the switch point machine installation site. In the control mode, many events are corresponding to the switch electric drive control mode: $\{s_1 - s_7, s_9, s_{12} - s_{16}\}$. To troubleshoot in control mode, on this set the question π_1 is asked: "Did the device failure at the signal tower?" Let us clarify the question for a wide range of readers – the elements of the control circuit are available both directly on the wayside (near the railway track) and in the relay-room of the signal tower. The set under consideration is divided into two subsets $\{s_{14}\}$ and $\{s_1 - s_7, s_9, s_{12} - s_{13}, s_{15} - s_{16}\}$. According to an outcome "1" of the issue, the event s_{14} is identified (the electric insulation violation of switch point machine). On the subset $\{s_1 - s_7, s_9, s_{12} - s_{13}, s_{15} - s_{16}\}$ the question π_2 is asked: "Is a malfunction detected after examining all parts of the switch point machine and control rulers?". The question separates an events subset into two multielement subsets. An outcome "1" subset of the question is separated by a new question π_3 ("Is a drive malfunction detected after a mechanical check of the control rulers misadjustment?"), And the outcome "0" subset by a new question π_4 ("A failure is detected in the device when measuring the insulation resistance and currents in the control circuits electric installation?"). Question π_3 identifies the following events: s_1 (the autotripper contact pads misadjustment), s_2 (the autotripper contact pads breakage), s_3 (the autotripper knife pads breakage), s_4 (the autotripper levers breakage), s_5 (the autotripper springs breakage), s_6 (the autotripper contact springs breakage), s_7 (the autotripper contact pads covered with hoarfrost), s_9 (the control rulers breakage), s_{12} (the slide gear mounting misadjustment). The π_4 identifies events s_{13} (hook-up wire breakdown) and s_{15} (the adjusting clip failure).

The Q^{I0}_D questionnaire is like the Q^T_D questionnaire with the only exception that the questions costs and the weight of the events are formed individually for each device.

The Q^{I1}_D questionnaire is built considering the possibilities of expanding the number of control points at which diagnostic parameters are measured automatically. Due to the increase in automatically controlled device parameters, new questions appear and the wording of existing one's changes. As an example of the reformulation of the question, we give the question π_3. In Q^{I0}_D, the question π_3 is formulated as follows: "Is a drive malfunction detected after a mechanical check of the following devices: loosening the control rulers or gate, gearbox malfunction?". In Q^{I1}_D, the question π_3 is formulated as follows: "Are there any extraneous shock-vibration effects?". Also added are questions π_7 ("Is the cable insulation resistance normal?") and π_8 ("Failures detected by the current transfer diagram and data from the accelerometer?"). Question π_7 is asked on the subset $\{s_{13} - s_{15}\}$. Event s_{13} is identified by an outcome "1" of this question, and events s_{14} and s_{15} are identified by an outcome "0". Question π_8 is asked on the subset $\{s_1, s_7, s_{10}, s_{12} - s_{15}, s_{23} - s_{24}, s_{26} - s_{28}\}$. By the outcome of "2" event s_{10}. is identified. Events s_{12} and s_{27}, are identified by an outcome "1" of the issue in question, and events s_{23}, s_{26}, s_{28} by "0". With the increase in control points, a diagnostic events subset expands, which includes events s_{24} (moisture), s_{25} (low oil level in the gearbox), s_{26} (non-smooth stroke of the switchblades), s_{27} (increased/decreased clearance between the switchblade and stock rail), s_{28} (probable contamination of rail shoe).

By analogy with the above for the operating mode, questions for the control mode are also added. Question π_5 is formulated as follows: "Is there a shock-vibration effect on the gate during the passage of the train?" Question π_6 is formulated as follows: "Is the humidity and temperature in the drive normal?".

D. Questionnaires for traffic lights

For the enters traffic light, the questionnaires shown in Fig. 8.

The Q_S questionnaire implies the use of question π_1, formulated as follows: "Is there a supply voltage in the transformer box?" As the root question. This question divides the original set of events into three subsets. Subsets of one and outcomes "0" are singleton. Question π_1 allows us to identify events s_6 (lower cable insulation) and s_7 (lack of power coming from the relay cabinet). The outcome "2" of the root question corresponds to the subset $\{s_1 - s_5\}$. On it, for further events identification, the question π_2: is asked: "Is there a standard voltage at the output from the signal transformer?" By the outcome "0" of this question, event s_1 (signal transformer failure) is identified. The subset of the single outcome of the question π_2 is multielement: $\{s_2 - s_5\}$. For its further breakdown, question π_3 ("What is the voltage at the lamp holder terminals?") Is used, and then – π_4 ("Is the voltage restored after cleaning the lamp holder?"). When posing the question π_3 it is possible to identify the events s_5 (lamp failure) and s_2 (open circuit of switching wires). When posing the question π_4 on a single outcome of the previous question, the events s_3 (poor contact in the lamp holder) and s_4 (switching damage) are identified.

The Q^T_D questionnaire for traffic lights is different from Q_S questionnaire. The initial events set in it is separate by the question π_1: "What is the type of failure?". Next, a subset of the one outcome of the root question is separate by the question π_3: "Is there a standard voltage at the output from the transformer box?" In this case, events s_5 (lamp failure) and s_1 (signal transformer failure) are detected. The subset of the outcome "0" of the root question is separate by the question π_2: "Is the supply voltage coming to the signal transformer?". This question can have three possible answers: outcome "2" – "Yes", outcome "1" – "No", outcome "0" – "Yes, but lowered". Question π_2 Is the supply voltage coming to the signal transformer?". This question can have three possible answers: outcome "2" – "Yes", outcome "1" – "No", outcome "0" – "Yes, but lowered". Question s_7 (lack of power coming from the relay cabinet) and s_6 (lower cable insulation resistance). There is also a third outcome (outcome "2") of the π_2, question, by which events s_2, s_3, s_4 are identified (their interpretation is given above).

The Q^{I0}_D questionnaire is like that described above, for which, however, the questions costs and the weight of the events are determined individually for each diagnostic object.

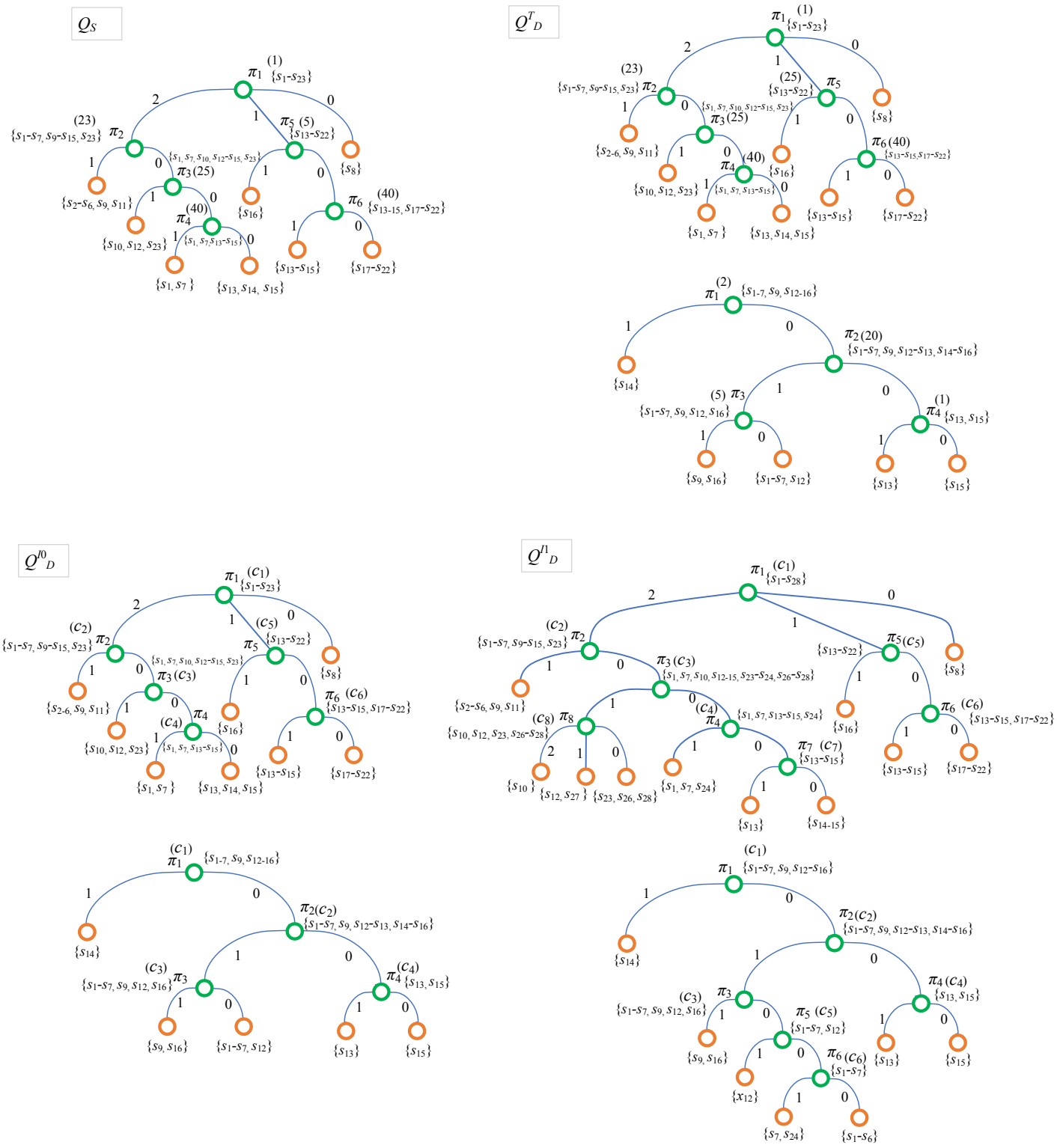


Fig. 7. Questionnaires family for switch point machines diagnostic.

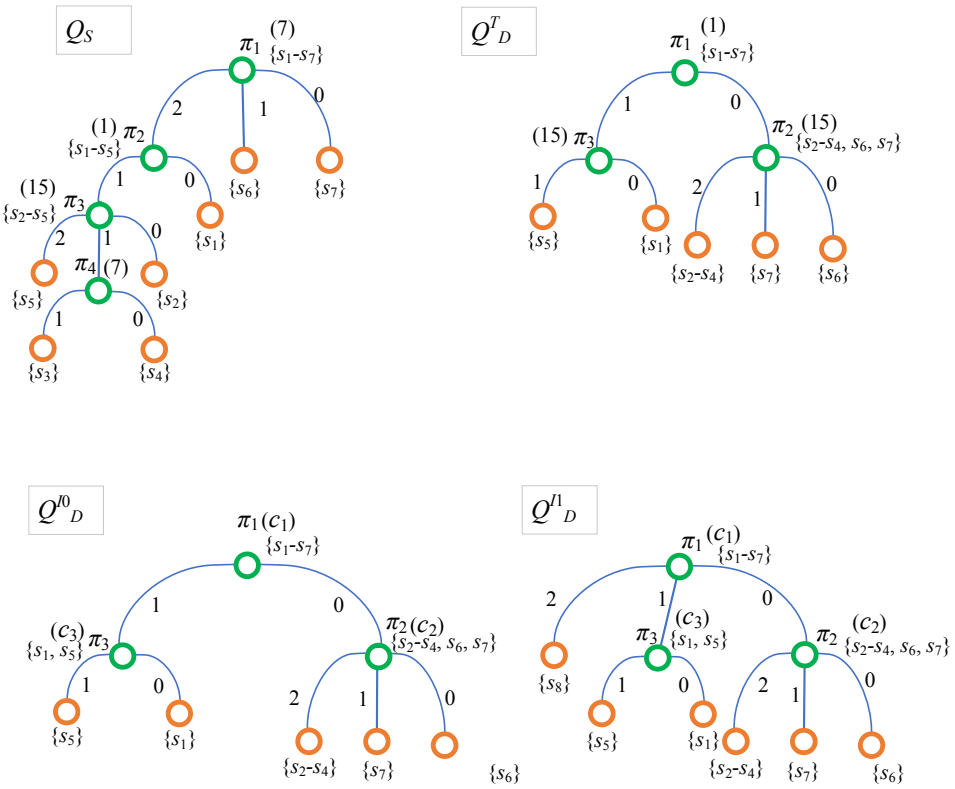


Fig. 8. Questionnaires family for traffic lights diagnostic.

The $Q^{I1}{}_D$ traffic light questionnaire may contain one more condition. For example, the addition of an inclinometer sensor allows the identification of an additional event s_8 (mast deviation). This event is identified by the third outcome of the root question π_1. It should be noted that in the diagnostic model of the traffic light (Fig. 5) there are several separate lamp switching circuits. Physically, these circuits are separated, but the power comes from a single source. In the given in Fig. 8 questionnaires, all the same type of faults for different traffic light indications are unified and combined. In other words, the questionnaire works for those lamps that are under load. All faults are identified for work objects in the corresponding mode.

In addition to the noted feature of traffic lights, it should also be added that the electric installation failure can be deepened to the level of one wire, since a bundle is placed in the mast of the traffic light, and separate wires feed various lamps in the bundle. In real conditions, the set of traffic light failures can be supplemented with such failure as a red light lamp failure, a red traffic light supply wire break, a yellow traffic light lamp failure, a yellow traffic light wire break, etc. This, however, can be considered a program-level monitoring system and highlight the failure logically.

IV. Algorithm for Formalizing the Initial Questions

Note the following feature of the proposed monitoring technology based on dynamic questionnaires. Initially, it is required to formalize the initial questions and add them to the database of the analytical subsystem of the monitoring system. Despite the apparent complexity of this task, it is not so difficult. All objects of railway automation can be classified into types and classes (like how this is done in computer-aided design systems for circuit solutions of railway automation [25]). The number of types and classes of railway automation devices is small. For example, the above three main types of wayside devices for railway automation are considered – railway track circuits, switch point machine and traffic lights. Classes are specific types of these devices – in this example, we used the class of audio-frequency railway track circuits, electric drives of the SP-6M brand with an AC motor and an enter five-signal traffic light. The monitoring system requires a description of the initial wording of the questions once for a specific type and class of railway automation devices. The algorithm for posing the initial questions is as follows:

1. Railway automation devices are divided into many types $T = \{T_1, T_2, \ldots, T_k\}$ and classes inside them

$$C^{T_i} = \left\{C_1^{T_i}, C_2^{T_i}, \ldots, C_{k_i}^{T_i}\right\}, C^{T_i} \in \{T_i\}.$$

2. For each type and class of devices, a state classifier is compiled (specific operational, operational, pre-failure, defective, defective, dangerous states):

$$S^{T_i,C^{T_i}} = \left\{S_1^{T_i,C^{T_i}}, S_2^{T_i,C^{T_i}}, \ldots, S_{q_i}^{T_i,C^{T_i}}\right\}.$$

3. Each state is associated with many diagnostic parameters with the permissible limits of their change to accurately identify the location of the object in this state.

4. For each specific automation object from a given type and class, the boundary values of the diagnostic parameters are set, for example, in automatic mode using self-adaptive networks (autoencoders) [26].

5. For each specific monitoring object, a valid question set is automatically selected according to the current operating mode, and arrays of question costs and event weights are generated. Initially, this is done based on general statistics

widely known for each of the devices. During the operation of the monitoring system for a specific automation object, the values of questions prices and event weights are adjusted considering the accumulated historical data about it.

Each state of any diagnostic object corresponds to a specific set of diagnostic features Δ_{S_r}, recorded by the monitoring system – the values of the diagnostic parameters in the aggregate, including taking into account the results of machine analysis and previous changes.

Confirmation 1. *For the technical diagnostics system and monitoring system to unambiguously interpret the obtained values set of diagnostic signs with any diagnostic object state, this set should not be its subset of diagnostic signs corresponding to another technical condition:*

$$\forall s_a: \Delta_{S_a} \not\subset \Delta_{S_x}, a \in \{1,2,\ldots,m\}, x \in \{1,2,\ldots,m\} \backslash \{a\}. \quad (1)$$

If condition (1) is not satisfied for at least one pair of technical conditions, then in automatic mode the monitoring system will not be able to determine the state in which the diagnostic object is located. This will require additional manual measurements. In dynamic questionnaires, such events will appear in one subset of the outcome of the question.

If it is impossible to separate any states, additional measurements are required by a certain algorithm.

Confirmation 2. *The states set* $S = \{s_1, s_2, \ldots, s_m\}$ *is separable if*

$$\forall s_i, i \in \{1,2,\ldots,r\}\ \exists \pi_j \in \Theta, j \in \{1,2,\ldots,m\}, \quad (2)$$

where π_j *– effective check.*

An effective check may follow from automatic measurement, or it may be the result of manual testing of a diagnostic object by service personnel. In this case, the monitoring system will give information messages to the operating personnel, facilitating the identification of the fault.

Thus, the process of formalizing the initial questions is implemented once for each type and class of railway automation devices using the expert method. Further, the monitoring system itself forms the admissible sequence of questions for each operating mode. The user of the monitoring system sees only the result of the search – the specific state of the device and the faulty component in case of failure or the recommended sequence of additional checks to localize the fault.

V. MONITORING SYSTEM ARCHITECTURE

Currently, monitoring systems in the field of railway transport are being built according to an architecture that excludes a remote cloud server for processing diagnostic information, and all processing consists only in comparing isolated measured parameters with a certain predefined thresh-old (normal) [16]. A more complex analysis of the parameters on the railways is used only "artificially" and has not been widely introduced. However, analysis methods, including those using artificial intelligence, are quite well developed, for example, in [27 – 29].

In Fig. 9 shows the perspective architecture of a monitoring system. In it, diagnostic data about all diagnostic parameters comes from measuring controllers with a set period for polling sensors in a data warehouse (hub). It also stores data on the physical characteristics and operating conditions of an object. According to a special protocol, a cloud server processing large amounts of data is accessed by the store and in real-time analyzes diagnostic data. In the case of fixing any malfunction (any violation of the regular operation of the devices), dynamic questionnaires are generated in the monitoring system software, from which, in turn, the recommended sequence of investigation of the incident (including the diagnostic algorithm) is automatically determined. To output data to the user, a virtual technological window is used (an analogue of a workstation), to which you can connect remotely from any mobile and stationary device connected to the Internet. Besides, the system can be supplemented by a stationary workstation, as well as mobile monitoring applications. This monitoring system architecture does not require the presence of central hubs, as in modern monitoring systems for railway automation devices [16]. Access to the monitoring results is carried out remotely through any browser or mobile application, regardless of the user's location.

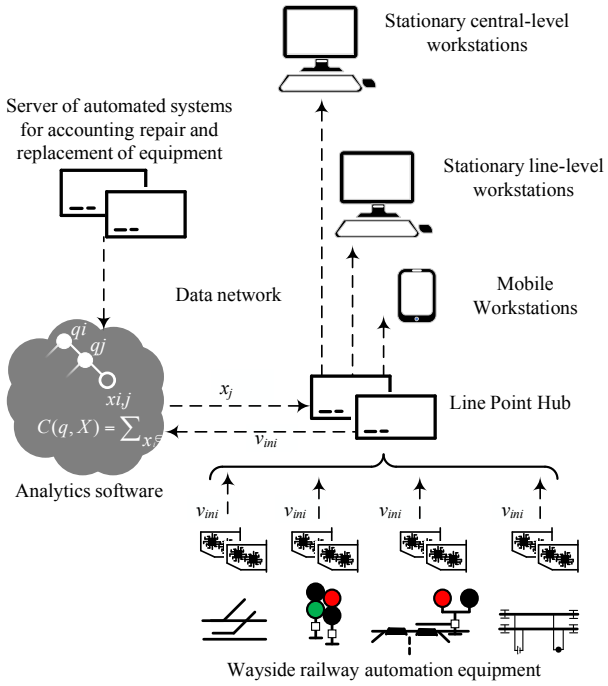


Fig. 9. Monitoring system architecture.

VI. CONCLUSION

Introduction to the dynamic questionnaires consideration allows approaching the process of constructing diagnostic algorithms for diagnosing modern systems with complex structures and many components much more accurately. Considering the work of monitoring systems in software that have access to large amounts of statistical data about the objects of diagnosis and the collected diagnostic data, it becomes possible to increase the completeness and depth of technical diagnosis. Also, the dynamic questionnaire itself in the monitoring system software forms the recommended sequence of investigation of the incident. This is the basis of decision support systems by technical personnel of operating organizations. The use of software tools with integrated dynamic questionnaires for automation devices thanks to cloud technology allows you to organize easy access to the monitoring results through any browsers and mobile applications. Access to the monitoring results is carried out using cyber-protected protocols, user identification and authentication

procedures according to the ranking of users and their settings.

The advantages of using dynamic questionnaires are that for any technical object it is possible to more accurately build a diagnostic algorithm taking into account both physical parameters and operating conditions, as well as accumulated statistical data on work and data from sensors of measuring controllers. It should be noted the disadvantages of using questionnaires for constructing diagnostic algorithms - the number of diagnostic events and questions for their identification significantly affects the speed of optimization of the questionnaire. Depending on the operating mode of the diagnostic object, the questionnaires may have different structures and bypass prices. Therefore, in real applications, it is necessary to consider the time limit for constructing the questionnaire, as well as, if necessary, apply the methods of decomposition or construction close to the optimal questionnaires.

In conclusion, we also note that the presented results are one of the applications of the theory of questionnaires. The use of dynamic questionnaires is also possible in other branches of science and technology.

REFERENCES

[1] P.P. Parkhomenko "Theory of Questionnaires (Review)", Automation and Remote Control, 1970, vol. 31, Issue 4, pp. 639-655.

[2] C.F. Picard "Graphs and Questionnaires", Netherlands: North-Holland Publishing Company, 1980, 431 p.

[3] P.P. Parkhomenko, and E.S. Sogomonyan "Technical Diagnosis Fundamentals (Diagnostic Algorithm Optimization, Apparatus Means)" (in Russ.), Moscow: Energoatomizdat, 1981, 320 p.

[4] P.P. Parkhomenko "Questionnaires and Organizational Hierarchies", Automation and Remote Control, 2010, Vol. 71, issue 6, pp. 1124-1134, doi: 10.1134/S0005117910060135.

[5] 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.

[6] A.Yu. Arzhenenko, and B.N. Chugaev "Optimal Binary Questionnaires" (in Russ.), Moscow: Energoatomizdat, 1989, 28 p.

[7] B.N. Chugaev, and A.Yu. Arzhenenko "Optimal Identification of Random Events" (in Russ.), Economics, Statistics and Informatics, 2013, Issue 2, pp. 188-190.

[8] 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.

[9] G. Duncan "Heterogeneous Questionnaire Theory", SIAM Journal on Applied Mathematics, 1974, Vol. 27, Issue 1, pp. 59-71, doi: 10.1137/0127005.

[10] 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.

[11] 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.

[12] D. Efanov, and V. Khoroshev "Method for Ordering Procedures of Dividing States by Procedures with Two and Three Results Taking into Account Their Cost and Weight of States" (in Russ.), SPIIRAS Proceedings, 2020, Vol. 19, Issue 1, pp. 218-243, doi: 10.15622/sp.2020.19.1.8.

[13] S.V. Mikoni, B.V. Sokolov, and R.M. Yusupov "Qualimetry of Models and Polymodel Complexes" (in Russ), Moscow, the Russian Academy of Sciences (RAS), 314 p.

[14] A.Yu. Arzhenenko, O.G. Kazakova, and V.A. Neyasov "Optimizing Binary Questionnaires with Variable Price Questions", Automation and Remote Control, 1989, Vol. 50, Issue 6, pp. 831-838.

[15] D.V. Efanov, and V.V. Khóroshev "Ternary Questionnaires", Proceedings of 17th IEEE East-West Design & Test Symposium (EWDTS`2019), Batumi, Georgia, September 13-16, 2019, pp. 289-300, doi: 10.1109/EWDTS.2019.8884404.

[16] 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.

[17] D.V. Efanov, G.V. Osadchy, V.V. Khóroshev, and D.A. Shestovitskiy "Diagnostics of Audio-Frequency Track Circuits in Continuous Monitoring Systems for Remote Control Devices: Some Aspects", Proceedings of 17th IEEE East-West Design & Test Symposium (EWDTS`2019), Batumi, Georgia, September 13-16, 2019, pp. 162-170, doi: 10.1109/EWDTS.2019.8884416.

[18] G. Theeg, and S. Vlasenko "Railway Signalling & Interlocking: 3ed Edition", Germany, Leverkusen PMC Media House GmbH, 2020, 552 p.

[19] 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.

[20] D. Sedykh, M. Gordon, D. Zuyev, and A. Skorokhodov "Analysis of the Amplitude and Phase-Manipulated Signals of Automation Devices via Bluetooth Technology", Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS`2018), Kazan, Russia, September 14-17, 2018, pp. 703-710, doi: 10.1109/EWDTS.2018.8524605.

[21] L. Heidmann "Smart Point Machines: Paving the Way for Predictive Maintenance", Signal+Draht, 2018 (110), issue 9, pp. 70-75.

[22] M. Wernet, M. Brunokowski, P. Witt, and T. Meiwald "Digital Tools for Relay Interlocking Diagnostics and Condition Assessment", Signal+Draht, 2019 (111), issue 11, pp. 39-45.

[23] D.V. Efanov "New Architecture of Monitoring Systems of Train Traffic Control Devices at Wayside Stations", Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS`2018), Kazan, Russia, September 14-17, 2018, pp. 276-280, doi: 10.1109/EWDTS.2018.8524788.

[24] 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.

[25] D. Sedykh, M. Gordon, and D. Efanov "Computer-Aided Design of Railway Signalling Systems in Russian Federation", Proceedings of 4th International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), Moscow, Russia, May 15-18, 2018, pp. 1-7, doi: 10.1109/ICIEAM.2018.8728662.

[26] Liou C.-Y., Cheng C.-W., Liou J.-W., and Liou D.-R. "Autoencoder for Words", Neurocomputing, 2014, Vol. 139, pp. 84-96, doi:10.1016/j.neucom.2013.09.055.

[27] W. Jin, Z. Shi, D. Siegel, P. Dersin, C. Douziech, M. Pugnaloni, P. La Cascia, and J. Lee "Development and Evaluation of Health Monitoring Techniques for Railway Point Machines", 2015 IEEE Conference on Prognostics and Health Management (PHM), 22-25 June 2015, Austin, TX, USA, DOI: 10.1109/ICPHM.2015.7245016.

[28] T. Böhm "Remaining Useful Life Prediction for Railway Switch Engines Using Artificial Neural Networks and Support Vector Machines", International Journal of Prognostics and Health Management 8 (Special Issue on Railways & Mass Transportation), December 2017, 15 p.

[29] T. Asada "Novel Condition Monitoring Techniques Applied to Improve the Dependability of Railway Point Machines", University of Birmingham, UK, Ph. D. thesis, May 2013, 149 p.

Optimizing Components of Multi-Module Systems Based on don't Care Input Sequences

Ekaterina Shirokova
Tomsk State University
Tomsk, Russia
k@shir.su

Larisa Evtushenko
Higher School of Economics
Moscow, Russia
evtlarisa@mail.ru

Andrey Laputenko
Higher School of Economics
Tomsk State University
Tomsk, Russia
laputenko.av@gmail.com

Nina Yevtushenko
Ivannikov Institute for System Programming of RAS
Higher School of Economics
Moscow, Russia
nyevtush@gmail.com

***Abstract*— In this paper, we use a window approach when optimizing Finite State Machine (FSM) components of a multi module system. Given a window with a loop-free binary composition of complete deterministic FSMs, we construct a partial FSM for the tail component FSM such that any reduced form of this partial FSM can replace the tail component preserving the composition behaviour. There are a number of cases when using a partial network equivalent instead of the initial component FSM allows to simplify the corresponding logic circuit with respect to the number of gates and path length between primary inputs and outputs.**

Keywords*— *Finite State Machine (FSM), Synchronous composition of FSMs, FSM equation component

I. Introduction

The Finite State Machine (FSM) model [1] is widely used to describe the behavior of systems that pass from state to state when an input is applied while producing output responses. Accordingly, FSMs are used in various fields to describe the behavior of control systems operating in the "request-response" mode and one of FSM based tasks is the system optimization, especially a component optimization with respect to various parameters. The Internet of Things (IoT) [2] can serve as an example, where various components are used that could be effectively optimized [2].

Optimization of multi-module systems is often iterative. In general case, when optimizing a component of an FSM composition, one can use the solution of a FSM equation of the form $A \bullet X \approx A \bullet B$ [3-5], where X is the component FSM under optimization and FSM A describes the joint behavior of other components and is often called the *context*. However, the complexity of solving such an equation is very high [4]. On the other hand, it is known [4, 5] that it is possible to select a "window" that contains a binary composition, and then optimize the components of this composition; in this case, the complexity of optimization will be significantly lower, especially if the window contains a loop-free composition.

The "window approach" in [5] uses "windows" which are more general and this provides additional difficulties when optimizing the component FSM. Also, in [5], the optimization with respect to the number of gates of the subsequent logic circuit is not discussed, while discussing the optimization with respect to other criteria such as the number of communication lines, fault tolerance with respect to certain faults, etc. In this paper, we discuss the optimization with respect to a subsequent logical circuit, that is, the optimization that later will be used for hardware implementations of FSM networks.

The work [6] describes the optimization of networks consisting of structural machines using internal "don't-care" sequences based on solving the feasibility problem for Boolean formulas; in our work, we rely on the construction of the network equivalent for the tail component according to the algorithm presented in [7]. The work [8] also touches the optimization topic and the authors consider not only input don't care sequences but also output don't care sequences for which the optimization is much more time-consuming. In [9], the optimization based on the solution of FSM equations was discussed and it was shown that sometimes it is more efficient to solve not one FSM equation but a system of such equations.

In this paper, we propose a procedure for solving an FSM equation for a binary loop-free FSM composition for a component for which all output channels are available for observation. This procedure is an adaptation of the general algorithm for solving FSM equations and generally speaking has polynomial complexity with respect to the size of the component FSMs. In this case, the general solution to the equation called also a network component equivalent, is a deterministic partial FSM which can be obtained in various ways. The use of a partial network equivalent instead of the initial component FSM in some cases makes it possible to simplify the corresponding logic circuit with respect to the number of gates, paths' length, etc, and this allows later on to optimize hardware FPGA implementations [10].

II. Preliminaries

A. Finite State Machines

An *initialized Finite State Machine* (FSM) S [4] or simply called an *FSM* or a *machine* throughout this paper is a 5-tuple $(S, I, O, s_0, \lambda_S)$, where S is a finite nonempty set of states with s_0 as the initial state; I and O are input and output alphabets; and $\lambda_S \subseteq S \times I \times O \times S$ is a transition relation. An initialized FSM can be considered as a function that maps sequences over one (input) alphabet into (subsets of) sequences over another (output) alphabet. If for each pair $(s, i) \in S \times I$ there exists at most one pair (o, s') such that $(i, s', o) \in \lambda_S$ then FSM S is *deterministic*; otherwise, FSM S is called *nondeterministic*. In

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

this paper, we consider deterministic FSMs if the converse is not explicitly stated. If for each pair $(s, i) \in S \times I$ there exists at least one pair (o, s') such that $(s, i, o, s') \in \lambda_S$ then FSM S is *complete*. Otherwise, FSM S is called *partial*. Given a state s of a partial FSM, for some input sequences, the behavior of the FSM at state s is not defined. The set $\Omega_S(s)$ is the set of input sequences under which the behavior of the FSM S at state s is defined.

States b and s of complete FSMs B and S are *equivalent* [11] if the FSM B at state b and the FSM S at state s produce equal (sets of) output sequences under any input sequence. Otherwise, the states b and s are called *distinguishable*. Initialized complete FSMs B and S are *equivalent*, written $B \approx S$, if their initial states are equivalent.

For partial FSMs, the definition of the quasi-equivalence is introduced [1, 11]. Given deterministic complete FSM B and possibly, partial FSM S, state b of B is *quasi-equivalent* to state s of a FSM S , written: $b \sqsupseteq s$, if $\Omega_B(b) \supseteq \Omega_S(s)$ and for each sequence $\alpha \in \Omega_S(s)$ the following holds: $out_S(s, \alpha) = out_B(b, \alpha)$. The initialized FSM B is *quasi-equivalent* to the initialized FSM S (notation: $B \sqsupseteq S$) if the initial state of the FSM B is quasi-equivalent to the initial state of the FSM S.

States s_i and s_j of the deterministic, possibly partial FSM S are *incompatible* or *distinguishable* if there exists an input sequence permissible for s_i and s_j such that the output responses to this sequence at states s_i and s_j do not coincide. The states s_i and s_j of the FSMS are *compatible* if they are not incompatible [11]. As an example of compatible states, states s_i and s_j can be considered where s_i is quasi-equivalent to s_j.

An initialized deterministic complete reduced FSM B is called a *reduced form* of an initialized deterministic possibly partial FSM S if B is quasi-equivalent to S. Differently from complete FSMs, a partial FSM can have several reduced forms which are not pair wise isomorphic. The reduced form of a FSM with a minimal number of states is called a *minimal form* of a FSM S. Again, a partial FSM can have several minimal forms which are not pair wise isomorphic.

B. Synchronous Composition of Finite State Machines

When designing complex systems, a system under design is usually represented as a composition of simpler subsystems and thus, composition operators over FSMs were defined in a number of publications [see, for example, 4, 12]. In this paper, we consider a so-called synchronous composition of initialized complete deterministic FSMs where each input is processed during one clock cycle. Such the composition is usually used to describe the composition of hardware components.

A synchronous composition of two FSMs (Figure 1) can be derived in various ways. In [12], a synchronous composition is constructed using the corresponding successor tree, however, the authors consider only the composition of Moore machines, i.e., the composition of complete deterministic FSMs where an output significantly depends only on a current state, that is, does not depend on an input [11]. In [4], it is proposed to build a synchronous composition using operations over languages of FSM components. The language of an FSM component is regular and can be represented by a finite automaton [13].

In this paper, we consider only loop-free composition and thus, use the composition definition based on a successor tree.

The successor tree describes the joint behavior of FSM components when applying external inputs of the set $I_1 \times I_2$ to the composition. The composition FSM is built from the successor tree by removing all internal transitions, i.e., each transition of the composition FSM is marked with an input symbol $i \in I_1 \times I_2$ and an external output symbol $o \in O_1 \times O_2$.

In this paper, a component FSM is optimized with respect to the behavior of other component FSMs. It is known [4] that the composition behavior not necessary changes when an FSM component is replaced by an FSM nonequivalent to the initial one. Usually, for such optimization, the explicit or implicit solving of the FSM equations is involved.

Given complete deterministic FSMs $A \bullet B = (S, I_1 \times I_2, O_1 \times O_2, s_0, \lambda)$ and $A = (A, I_1, O_1 \times U, a_0, \lambda_A)$ (Figure 1), the expression $A \bullet X \approx A \bullet B$, in which X is a FSM over an input alphabet $I_2 \times U$ and an output alphabet O_2, is called a *synchronous FSM equation*.

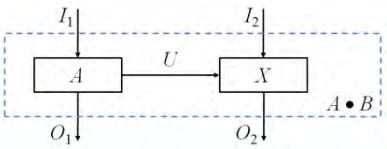


Fig. 1. The synchronous composition of FSMs

A FSM B' with an input alphabet $I_2 \times U$ and an output alphabet O_2 is called a *solution* of the equation $A \bullet X \approx A \bullet B$ if $A \bullet B' \approx A \bullet B$. A solution of the FSM equation $A \bullet X \approx A \bullet B$ corresponds to the FSM, which, when combined with the FSM A, determines the behavior that coincides with the behavior of the FSM $A \bullet B$. The solution *Largest* is called the largest solution to the equation $A \bullet X \approx A \bullet B$ if each solution to the equation is a reduction of the FSM *Largest* and it is known that such the largest solution exists [4].

III. COMPONENT-ORIENTED OPTIMIZATION OF A SYNCHRONOUS FSM COMPOSITION

We propose an algorithm for solving the FSM equation for a binary loop-free FSM composition, when all output channels of a component FSM under optimization are available for observation.

A. Optimizing the tail component of a serial binary composition

Consider a serial composition of complete deterministic FSMs A and B (Figure 2). To derive an FSM that describes the behavior of the component B that affects the composition behavior, i.e., a so-called *network equivalent* of B, we propose two procedures. We first derive the so-called *inverse automaton* A^* of the component A by leaving only outputs at each transition; if necessary, we determinize the obtained automaton and intersect this automaton with B [7]. The obtained possibly partial machine is the network equivalent of B. Input sequences and only those where the

behavior of B can change the composition behavior are defined for the network equivalent.

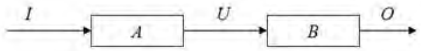


Fig. 2. The serial composition of FSMs A and B

Proposition 1 [7]. A complete FSM C can replace the component B preserving the external composition behavior if and only if C is quasi-equivalent to the network equivalent of B.

Consider an example from [10] where A and B are component FSMs of a serial composition (Figure 2). Figure 3(c) shows the inverse automaton A^* for the component A.

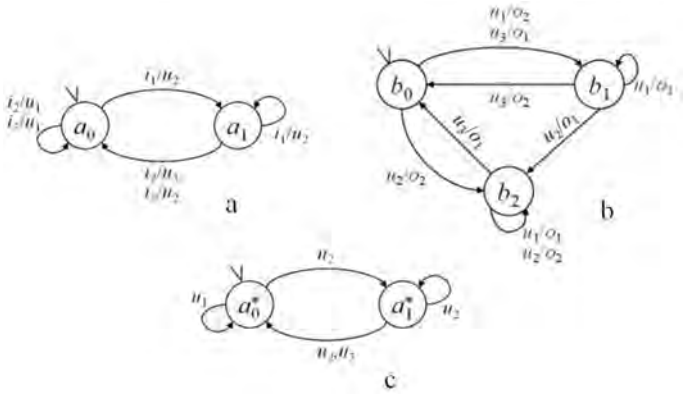


Fig. 3. (a) The FSM A; (b) The FSM B, (c) The inverse automaton A^*

After the determinization, the states of the inverse automaton can become subsets of states of the initial head component FSM and at the next step, the network equivalent of B is derived as the product of the inverse automaton and the initial tail component FSM. The network equivalent is the largest solution to the equation and according to the above procedure, the network equivalent can have more states than the initial tail component FSM. However, according to our experiments, usually the network equivalent has less transitions than the initial tail component FSM.

In order to keep the number of states in a solution equal to the number of B states, a part of the network equivalent can be derived. For this purpose, a serial composition of two component FSMs over all composition actions is constructed based of the component successor trees. Given a state b of the tail component, a transition at state b under input u is left undefined if and only is there is no state (a_j, b) in the composition such that the output u can be produced at state a_j; otherwise, the transition of B is left intact; such a machine is denoted $Comp(B)$.

Proposition 2. If a complete FSM C is quasi-equivalent to the FSM $Comp(B)$ then C can replace the component B preserving the external composition behavior.

For the component FSMs in Figure 3, the composition successor tree and the composed FSM are shown in Figures 4(b) and 4(c). The network equivalent B' is the same as the machine $Comp(B)$ (Figure 4(a)).

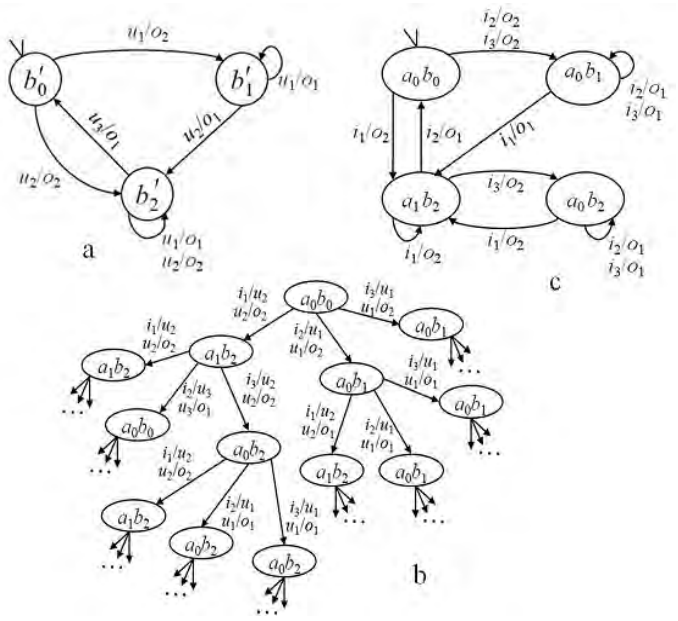


Fig. 4. (a) The machine B' (The machine $Comp(B)$), (b) The successor tree (c) The composed FSM

Consider now how the tail component FSM can be optimized in a loop-free binary composition with a more complex structure.

B. Optimizing the tail component of a binary loop-free composition

If the composition is loop-free but the tail component has additional input channels then these channels should be taken into account when deriving the network equivalent for the tail component. Consider various cases of such composition.

Case 1. Figure 5(a) shows the composition of FSMs A and B. The FSM A (Figure 5(b)) has an input alphabet I and an output alphabet U while the FSM B has an input alphabet $I \times U$.

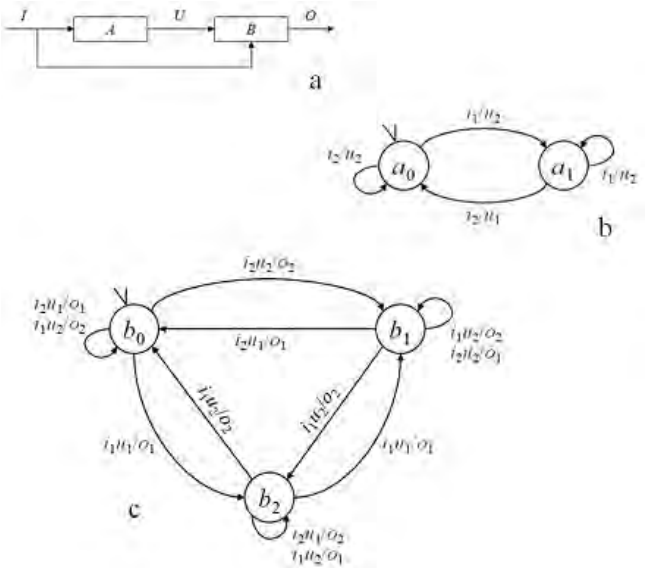


Fig. 5. (a) The synchronous composition of A and B; (b) The FSM A; (c) The FSM B

In this case, the inverse automaton is derived over the Cartesian product $I \times U$ and then the network equivalent of the tail component FSM for the machine $Comp(B)$ is derived as described above.

In Figure 5, $I = \{i_1, i_2\}$ and alphabet $U = \{u_1, u_2\}$.The FSM B (Figure 5(c)) has input alphabet $I \times U$ and an output alphabet $O = \{o_1, o_2\}$.

In Figure 6, there are the automaton A^* and the reduced form of the network equivalent B'. The network equivalent B' has fewer transitions and less states than the initial FSM B .

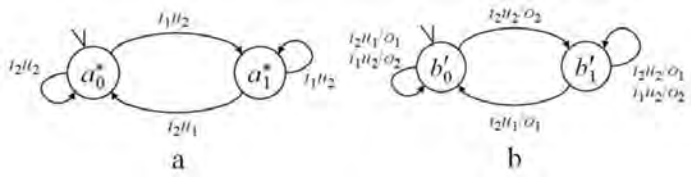


Fig. 6. (a) The automaton A^*; (b) The network equivalent B'

Case 2. Consider another loop-free composition adding an input channel with alphabet V to the component B (Figure 7(a)). Extend each transition in the FSM B to the alphabet V: at each transition add the symbol "–", which means that there can be any symbol of the alphabet V.

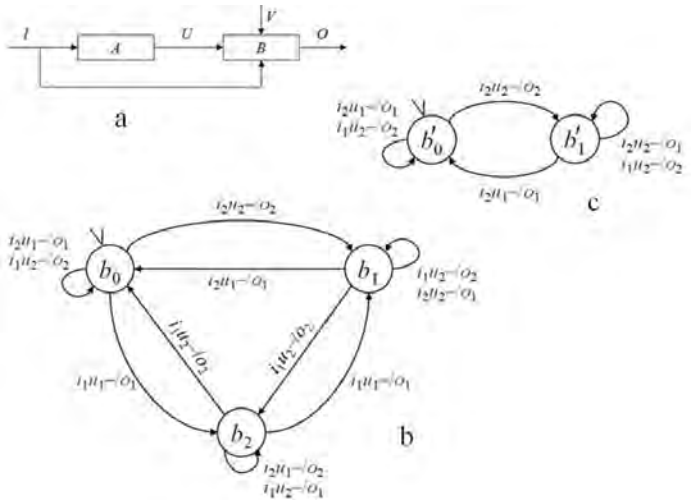


Fig. 7. (a) The synchronous composition of A and B; (b) The FSM B; (c) The network equivalent B'

In Figure 7, $V = \{v_1, v_2\}$ and the FSM B (Figure 5(c)) extension is shown in Figure 7(b). Thus, to obtain the network equivalent, it is sufficient to expand each transition of the FSM B' (Figure 6(b)) to the alphabet V. Figure 7(c) shows the network equivalent.

We can also consider a more complex network in the window. Consider an example of a network of three FSMs (Figure 8). In the network, we can select a "window" with a binary loop-free (or even serial) FSM composition of components A and B, and for the tail component of a such composition it is enough to simply calculate the working area of operating: the component B receives inputs of the alphabet V from the component A and of the alphabet U from the component C; component FSM B responses to this symbols by outputs of the alphabet O. Thus, the behavior of component B is essential for the external behavior of the network only on those input sequences that can appear at the outputs of component FSMs C and A. All other input sequences are so called don't care input sequences for the component B where this FSM component can have any behavior.

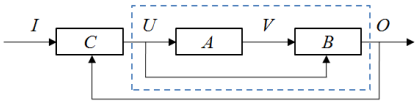


Fig. 8. The network of three FSMs

IV. Implementation of Synchronous Composition as FPGA

Various solutions can be proposed for representing the network equivalent as a logic circuit which later on will be implemented using the FPGA technology. Given an FSM, in order to derive a logic circuit, FSM states, inputs and outputs symbols are encoded as Boolean vectors. The network equivalent of B can be minimized first and a logic circuit is derived for an obtained complete reduced FSM. On the other hand, representing the network equivalent as a logic circuit can be done without minimization as sometimes optimal logic circuits can be obtained for non-minimal FSMs.

Another approach is to implement not the network equivalent of B but the machine $Comp(B)$ and this also can be used for optimization of an FSM B logic circuit representation .

Using a partial network equivalent instead of the initial component allows in some cases to simplify the corresponding logic circuit with respect to the number of gates and path length from external input ports to external output ports. As an example, consider a sequential composition of two combinational circuits in Figure 9(a) [3]. The head circuit consists of two disjunctors, and the tail circuit consists of a conjunctor and an inverter. Accordingly, only sets 00 and 11 can be obtained from the outputs of the head component. At the input set 00, the tail component "responds" with output symbol 1; the tail component "responds" to the output set 11 with the output symbol 0. Thus, the tail component can simply be replaced by an inverter, the input of which is the output of one of the disjunctors (Figure 9(b)). However, when using combinational circuits after the optimization, the whole system should be checked for not having combinational loops.

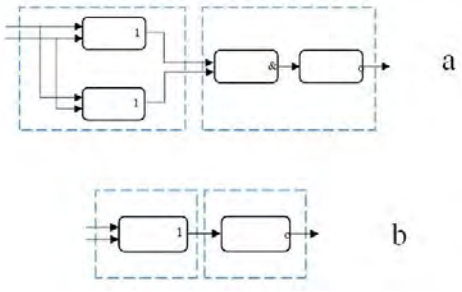


Fig. 9. An example of optimizing a serial network of two combinational circuits

Usually, the more undefined the finite state machines are, the simpler the corresponding logic circuits can be obtained. Correspondingly, we performed a preliminary estimation for the number of undefined transitions in the resulting tail component network equivalent.

Proposition 3. Let complete FSMs *A* and *B* be then head and tail components of the composition. If each output is possible at each state of the FSM *A*, then the network equivalent of *B* is equivalent to *B* .

However, if approximately half of the outputs are defined at each state of the head component, then about 30% of undefined transitions appear in the network equivalent.

As a future work, we are going to implement FSM components and their network equivalents using FPGA, since again according to our preliminary experiments, a large number of undefined transitions allows to reduce the maximum length of a path from terminal inputs to outputs in the logic circuit, i.e., increase the frequency of the hardware implementation derived by FPGA technology.

REFERENCES

[1] N.V. Yevtushenko, A.F. Petrenko, M.V. Vetrova, 2006 "Nondeterministic finite state machines: analysis and synthesis. Part 1. Relations and operations," Tomsk, Tomsk State University, p. 142, 2006. (In Russian).

[2] Xia F., Yang L.T., Wang L., Vinel A., "Internet of Things," International Journal of Communication Systems, Volume 25, No. 9, pp 1101-1102, 2012.

[3] N.V. Yevtushenko, M.V. Rekun and S.V. Tihomirova, "Non-deterministic FSMs: analysis and synthesis. Part 2. Solving FSM Equations," Tomsk, Tomsk State University, p. 111, 2009. (In Russian).

[4] T. Villa, N. Yevtushenko, R.K. Braytuon, A. Mishchenko, A. Petrenko, A.L. Sangiovanni-Vincentelli, "The Problem of the Unknown Component: from Theory to Applications," Springer, p. 323, 2011.

[5] S.V. Tikhomirova, "Optimization of multicomponent discrete systems based on the solution of FSM equations," The dissertation for the degree of candidate of technical sciences, Tomsk, Tomsk State University, 2008. (In Russian).

[6] A. Mishchenko, R.K. Brayton, "SAT-based complete don't-care computation for network optimization," The Proceedings of the Design, Automation and Test in Europe Conference, Volume 01, pp. 412-417, March 2005.

[7] M.V. Vetrova, "Development of synthesis and testing algorithms for finite-automaton compensators," The dissertation for the degree of candidate of technical sciences, Tomsk, Tomsk State University, 2003. (In Russian).

[8] J.-K. Rho, F. Somenzi, Don't care sequences and the optimization of interacting finite state machines," IEEE Trans. Comp. Aided Des., Volume 13, No. 7, pp. 865-874, 1994.

[9] N. Yevtushenko, S. Zharikova, M. Vetrova, "Multi component digital circuit optimization by solving FSM equations," Proceedings of the Euromicro Symposium on Digital Systems Design, DSD '03, pp. 62–68, 2003.

[10] V.A. Schwarzkop, "Optimization of components of multimodular systems based on the solution of automaton equations," Master's thesis, Tomsk, Tomsk State University, 2019. (In Russian).

[11] A. Gill, "Introduction to the theory of finite-state machines," Moscow, Nauka, p. 272, 1966.

[12] J. Hartmanis, R.E. Stearns, "Algebraic Structure Theory of Sequential Machines," Prentice-Hall, 219 p, 1966.

[13] J.E. Hopcroft, R. Motwani, J.D. Ullman, "Introduction to automata theory, languages, and computation," Addison-Wesley, 521 p, 2001.

Sampling Theorem in Time Domain for Infinite Duration Signal: Analytical Expression and Geometric Illustration

Gamlet S. Khanyan
Central Institute of Aviation Motors
Moscow, Russia
khanyan@rtc.ciam.ru

***Abstract*—The work is devoted to the study of sampling theorem for a signal of infinite duration by two complementary methods – analytical and geometric ones. A comparison is made with the results obtained previously for a signal of finite duration.**

Keywords—harmonic signal, frequency band index, canonical variables, summation of series

I. INTRODUCTION

One of the fundamental provisions of information theory is the time domain sampling theorem widely known in the Whittaker-Kotelnikov-Shannon classical formulation [1] as a method for reconstructing from its readings $s(t_n)$ obtained with sampling rate F the infinite duration signal $s(t)$ determined by a spectral function $S(f)$ of limited frequency band $0 \le f < F$, so that both spectra of amplitudes $A(f) = 2|S(f)|$ and phases $\Phi(f) = \arg S(f)$ are restricted by upper cut-off frequency $F/2$.

Less well-known are versions of the theorem with a finite number of samples N (see, for example, [2], where N is odd) and with bandpass filtering where the cut-off frequencies $F_{low} = GF/2$, $F_{up} = (G+1)F/2$ are determined by the frequency band index $G \ge 0$ (of integer value – to avoid frequency overlapping – see, for example, [3]-[4]). These and other numerous generalizations of the theorem (non-trigonometric kernel, non-uniform sampling, cutting errors, noise, etc.) are covered in extensive literature presented with sufficient completeness (248 sources) in a survey article cited by many authors [5].

The author's research ([6]-[8], etc.) is devoted to bandpass filtering with infinite and arbitrary finite sample number N of a signal observed in frequency band with an arbitrary real number index of G. One of the initial results of theorem's infinite version – the formulas for cut-off frequencies (37) – was established in [6] proceeding from the idea that for the absence of frequency overlay the periodic continuation $S(f+gF)$ should not intersect with the mirror reflections $A(-f)$ and $\Phi(-f)$, where, as it turned out, $g = [G]$, which was confirmed in [7]-[8] by analyzing the transform (42) of the base signal – a fragment of harmonic oscillations (1) of duration $T = N/F$.

In the present paper, a comprehensive analysis of the infinite version of theorem [1], started in [6], is carried out.

II. THE SUBJECT OF STUDY. PURPOSE OF THE WORK

A harmonic signal of infinite duration

$$s(t) = a\cos(2\pi f t + \varphi), \quad -\infty < t < +\infty \tag{1}$$

with parameters $a \ge 0$, $-\infty < f < +\infty$ and $-\pi < \varphi \le \pi$ is subjected at a base time instant t_0 to discretization with a frequency F:

$$s(t_n) = \int_{-\infty}^{+\infty} s(t)\delta(t - t_n)dt = a\cos(2\pi f t_n + \varphi);\ t_n = t_0 + n/F, \tag{2}$$

after which it is restored by the transform

$$s'(t) = \sum_{n=-\infty}^{+\infty} s(t_n)\frac{\sin\pi(G+1)F(t-t_n) - \sin\pi GF(t-t_n)}{\pi F(t-t_n)} \tag{3}$$

the kernel of which contains three parameters – sample rate F, band index G, and base time t_0.

The conversion result, $s'(t)$, is marked by a stroke since the conditions imposed on the kernel parameters F, G, t_0 and on the signal ones a, f, φ, under which the transform (3) is identical:

$$s'(t) = s(t), \tag{4}$$

are far from obvious, which makes up the content of sampling theorem for the infinite duration harmonic signal: signal (1) is recovered by its samples (2) via interpolation formula (3).

As soon as the frequency band boundaries F_{low}, F_{up} within which (4) takes place are found somehow dependent on parameters listed above, then, due to linearity of transform (3), the time domain sampling theorem will be valid for the signal

$$s(t) = \int_{F_{low}}^{F_{up}} a(f)\cos(2\pi f t + \varphi(f))\,df \tag{5}$$

representing a superposition of harmonic oscillations (1).

The purpose of this work is to define these boundaries, as well as to describe the phenomena that occur both when the theorem is true and when it is violated.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

III. Transition to Dimensionless Quantities

We pass in (1)-(2) to the current time $q = Ft$ and the signal frequency $p = 2f/F$ undimensioned over parameter F, as a result of which (3) takes the form

$$s'(t) = a\sum_{n=-\infty}^{+\infty}\cos(\pi p q_n + \varphi)\frac{\sin\pi(G+1)(q-q_n) - \sin\pi G(q-q_n)}{\pi(q-q_n)} \quad (6)$$

where $q_n = Ft_n = q_0 + n$ is the sequence of dimensionless time q samples with the base value $q_0 = Ft_0$.

Let us introduce instead of frequency p and index G a couple of dimensionless parameters associated with them:

$$\xi = \frac{G+p+1}{2}, \quad \eta = \frac{G-p+1}{2}, \quad (7)$$

which we call canonical variables in the sense (as will be explained later) that they represent p and G in an equal form on the (p, G) plane – by analogy with the equal status of generalized coordinates and momenta in the Hamilton function of a physical system in analytical mechanics (or their operators in quantum mechanics).

The transform (6), considered from now as a function of variables (ξ, η), takes the form:

$$s'(t) = a\sum_{n=-\infty}^{+\infty}\cos(\pi(\xi-\eta)(q_0+n)+\varphi)\times \quad (8)$$
$$\times\frac{\sin\pi(\xi+\eta)(q-q_0-n) - \sin\pi(\xi+\eta-1)(q-q_0-n)}{\pi(q-q_0-n)}.$$

IV. Calculating the Transform

Representing the product of cosine (signal's digital realization) on each of sines in the numerator of transform (8) kernel as the sum of two sines and changing the summation index n by $-n$, we get a four-term expression

$$s'(t) = W(\xi,\eta,\varphi) + W(\eta,\xi,-\varphi) - \quad (9)$$
$$- W(\xi-1/2,\eta-1/2,\varphi) - W(\eta-1/2,\xi-1/2,-\varphi)$$

with basic (mother) function

$$W(\xi,\eta,\varphi) = a\sum_{n=-\infty}^{+\infty}\frac{\sin(\pi(\xi+\eta)q - \varphi - 2\pi\xi q_0 + 2\pi\xi n)}{2\pi(q-q_0+n)}. \quad (10)$$

Disclosure of sine in the numerator turns the right-hand side (10) into a binomial expression

$$W(\xi,\eta,\varphi) = \frac{a}{2}U(\xi)\sin(\pi(\xi+\eta)q - \varphi - 2\pi\xi q_0) +$$
$$+\frac{a}{2}V(\xi)\cos(\pi(\xi+\eta)q - \varphi - 2\pi\xi q_0) \quad (11)$$

where $U(\xi)$ and $V(\xi)$ are trigonometric series whose sums (for non-integer $q - q_0$) are tabulated in the extensive reference book on integrals and series [9], item 5.4.3.4:

$$U(\xi) = \sum_{n=-\infty}^{+\infty}\frac{\cos 2\pi\xi n}{\pi(q-q_0+n)} = \frac{\cos\pi(q-q_0)(1-2\{\xi\})}{\sin\pi(q-q_0)}, \quad (12)$$
$$V(\xi) = \sum_{n=-\infty}^{+\infty}\frac{\sin 2\pi\xi n}{\pi(q-q_0+n)} = \frac{\sin\pi(q-q_0)(1-2\{\xi\})}{\sin\pi(q-q_0)}\operatorname{sgn}\{\xi\}.$$

The second function, $V(\xi)$, for the integer and half-integer values of variable ξ is equal to zero (when the sign of its fractional part $\{\xi\}$ is zero and when the fractional part itself is 1/2 for its sign equal to unity). The same function, for this reason, is equal to zero for the remaining integer and half-integer values of its argument in (9): ξ–1/2, η, η–1/2.

Consider $V(\xi - j/2) \neq 0$ and $V(\eta - j/2) \neq 0$, where $j = 0$ or 1, as the main case, and the equality of these functions to zero as the special case of a pair of parameters (ξ, η) and their prototypes (p, G). These two cases can be distinguished by whether the expressions $\operatorname{sgn}\{2\xi\}$ and $\operatorname{sgn}\{2\eta\}$ are equal to unity or zero, respectively. In the latter case, the variables ξ, η will, for the same values of j, be represented as $\xi = k + j/2$, $\eta = l + j/2$; $k, l = 0, \pm1, \pm2, \ldots$.

A. The Main Case

Since $\xi \neq k + j/2$, we have $\operatorname{sgn}\{\xi\} = \operatorname{sgn}\{2\xi\} = 1$ in (12), then

$$W(\xi,\eta,\varphi) = a\frac{\cos u\sin v + \sin u\cos v}{2\sin\pi(q-q_0)}; \quad (13)$$
$$u = \pi(1-2\{\xi\})(q-q_0), \quad v = \pi(\xi+\eta)q - \varphi - 2\pi\xi q_0.$$

Using the formula of sine of sum of two angles we express two of the functions (9) in explicit form:

$$\begin{cases} W(\xi,\eta,\varphi) = a\dfrac{\sin(u+v)}{2\sin\pi(q-q_0)} = a\dfrac{\sin(\pi(\eta-\xi)q + \pi(1+2[\xi])(q-q_0) - \varphi)}{2\sin\pi(q-q_0)} \\ W(\xi-1/2,\eta-1/2,\varphi) = a\dfrac{\sin(\pi(\eta-\xi)q + \pi(1+2[\xi-1/2])(q-q_0) - \varphi)}{2\sin\pi(q-q_0)} \end{cases}. \quad (14)$$

We calculate the difference between these functions by the formula of difference of sines of two angles using the basic properties $[\xi + k] = [\xi] + k$, $\{\xi + k\} = \{\xi\}$; $k = 0, \pm1, \pm2, \ldots$ of operations for taking integer and fractional parts of numbers:

$$s_{\xi,\eta}(\varphi) = W(\xi,\eta,\varphi) - W(\xi-1/2,\eta-1/2,\varphi) =$$
$$= -a\frac{\sin\pi[\{\xi\}-1/2](q-q_0)}{\sin\pi(q-q_0)}\times \quad (15)$$
$$\times\cos(\pi(\eta-\xi)q + \pi(2[\xi] + [\{\xi\}+1/2])(q-q_0) - \varphi).$$

It turns out that the multiplier $j = [\{\xi\} - 1/2]$ for $(q - q_0)$ in sub-sine expression of fraction's numerator in (15) is 0 or -1. In fact, this equality, by definition of integer part, is equivalent to the inequality $j \le \{\xi\} - 1/2 < j + 1$, which, it is easy to verify, will be compatible with inequality $0 \le \{\xi\} < 1$ defining fractional part, in two mutually exclusive cases: $1/2 \le \{\xi\} < 1$ (for $j = 0$) and $0 \le \{\xi\} < 1/2$ (for $j = -1$). In the latter case, the term $[\{\xi\} + 1/2]$ in sub-cosine expression equals to zero due to the inequality $0 \le 1/2 \le \{\xi\} + 1/2 < 1$, therefore, it can be ignored not only for $j = -1$, but also for $j = 0$, when, due to equality of the whole fraction to zero, the entire right-hand side (15) turns out to be zero. Noting that for $j = -1$ the sines in the numerator and the denominator are reduced, the final stage of simplification (15) is presented as:

$$s_{\xi,\eta}(\varphi) = a(-[\{\xi\} - 1/2])\cos(\pi(\xi-\eta)q - 2\pi[\xi](q-q_0) + \varphi). \quad (16)$$

In the same way, we calculate the difference between the two signal components remaining in (9):

$$s_{\eta,\xi}(-\varphi) = W(\eta,\Xi,-\varphi) - W(\eta-1/2,\Xi-1/2,-\varphi) = \\ = a(-[\{\eta\}-1/2])\cos(\pi(\eta-\Xi)q - 2\pi[\eta](q-q_0)-\varphi) \quad (17)$$

The converted signal (9), in final view, is the sum of two components (16) and (17):

$$s'(t) = s_{\Xi,\eta}(\varphi) + s_{\eta,\xi}(-\varphi). \quad (18)$$

B. The Special Case

Let us first consider the case $\operatorname{sgn}\{2\Xi\}=0$ without imposing any restrictions on η. Then $\Xi = k + j/2$ for j and k specified above. Using formulas (11) - (15), having there $V(\Xi) = V(\Xi - 1/2) = 0$, we get the following expression for the first component of signal (18):

$$s_{\Xi,\eta}(\varphi) = a\frac{U(\Xi)\sin v - U(\Xi-1/2)\sin(v-w)}{2}, \quad w = \pi(q-q_0). \quad (19)$$

Using the above properties of integer and fractional part operations, we establish that

$$U(\Xi) = \frac{\cos(1-2\{k+j/2\})w}{\sin w} = \frac{j+(1-j)\cos w}{\sin w}, \quad (20)$$
$$U(\Xi-1/2) = \frac{\cos(1-2\{k+j/2-1/2\})w}{\sin w} = \frac{(1-j)+j\cos w}{\sin w}.$$

Substituting (20) into (19) and making equivalent trigonometric transformations, we obtain:

$$s_{\Xi,\eta}(\varphi) = a\frac{j+(1-j)\cos w}{2\sin w}\sin v - a\frac{(1-j)+j\cos w}{2\sin w}\times \\ \times(\sin v\cos w - \cos v\sin w) = a\frac{\cos(v-jw)}{2}. \quad (21)$$

Now we substitute v and w from (13) and (19) into (21), taking into account that $j = 2(j/2) = 2(\Xi - k) = 2(\Xi - [\Xi])$:

$$s_{\Xi,\eta}(\varphi) = \frac{a}{2}\cos(\pi(\Xi-\eta)q - 2\pi[\Xi](q-q_0)+\varphi). \quad (22)$$

In the same way, we calculate the second component of the signal (18) for the integer and half-integer values of η:

$$s_{\eta,\xi}(-\varphi) = \frac{a}{2}\cos(\pi(\eta-\Xi)q - 2\pi[\eta](q-q_0)-\varphi). \quad (23)$$

C. Combining the Main and Special Cases

Comparing expressions (22) and (23) with expressions (16) and (17) we notice that they differ only in coefficients at cosines that are identical for each of the components. As was established above, the amplitude of Ξ-component $s_{\Xi,\eta}(\varphi)$ of transformed signal (18) in the main case of (16) is a if $0 < \{\Xi\} < 1/2$, and 0 if $1/2 < \{\Xi\} < 1$, while in the special case (22), when $\{\Xi\} = 0$ or $\{\Xi\} = 1/2$, this amplitude is $a/2$. One can combine both cases as follows:

$$s_{\Xi,\eta}(\varphi) = a\frac{1+(-1)^{[2\Xi]}\operatorname{sgn}\{2\Xi\}}{2}\times \\ \times\cos(\pi(\Xi-\eta) - 2\pi[\Xi](q-q_0)+\varphi). \quad (24)$$

It can be seen that if $\operatorname{sgn}\{2\Xi\}=0$, then the numerator of the fraction before the cosine is 1, and the amplitude is $a/2$; if $\operatorname{sgn}\{2\Xi\}=1$, then this numerator is 2 or 0, depending on whether the integer $[2\Xi]$ is even or odd. If it is even (equal to $2k$), then the definition of integer part $2k \le 2\Xi < 2k+1$ implies $k < \Xi < k+1/2$, which means $0 < \{\Xi\} < 1/2$; if odd, then $2k+1 \le 2\Xi < 2k+2$ implies $1/2 < \{\Xi\} < 1$. In the same way, changing the sign of sub-cosine expression, we generalize the η-component of the signal:

$$s_{\eta,\xi}(-\varphi) = (a/2)(1+(-1)^{[2\eta]}\operatorname{sgn}\{2\eta\})\times \\ \times\cos(\pi(\Xi-\eta) + 2\pi[\eta](q-q_0)+\varphi). \quad (25)$$

D. Converting the Nodal Points

This is what we call the current time counts $t = t_m$; $m = 0, \pm 1, \pm 2, \ldots$ taken in (3) at the signal samples. In this case $q - q_0 = m$ and formulas (12) are not applicable (the denominator $\sin\pi m$ in the expressions for U and V becomes zero). Therefore, we calculate transform (8) by another method. Substituting there $q_0 = q - m$, we have:

$$s'(t) = a\sum_{n=-\infty}^{+\infty}\cos(\pi(\Xi-\eta)(q-m+n)+\varphi)\times \\ \times\frac{\sin\pi(\Xi+\eta)(m-n) - \sin\pi(\Xi+\eta-1)(m-n)}{\pi(m-n)}. \quad (26)$$

Replacing $n - m$ with the summation index n and applying the formula for the difference of sines of two angles we bring (26) to a form that does not contain m explicitly:

$$s'(t) = a\sum_{n=-\infty}^{+\infty}\cos(\pi(\Xi-\eta)(q+n)+\varphi)\cos\pi(\Xi+\eta-1/2)n\frac{\sin\pi n/2}{\pi n/2}. \quad (27)$$

The main problem – division by zero in the kernel of transform (27) – is solved by isolating the summand of series with number $n = 0$, using for this the first remarkable limit $(\sin x)/x \to 1$ when $x \to 0$ (where $x = \pi n/2$) and combining the sums over positive and negative n:

$$s'(t) = a\cos(\pi(\Xi-\eta)q+\varphi) + \\ + a\sum_{n=1}^{+\infty}\begin{pmatrix}\cos(\pi(\Xi-\eta)(q+n)+\varphi)+\\ +\cos(\pi(\Xi-\eta)(q-n)+\varphi)\end{pmatrix}\cos\pi(\Xi+\eta-1/2)n\frac{\sin\pi n/2}{\pi n/2}. \quad (28)$$

The terms of series with even numbers $n = 2k$ are equal to zero (due to $\sin\pi k = 0$ when $\pi k \ne 0$); whereas for odd numbers $n = 2k+1$ formula (28) after summing the cosines in large parentheses takes the form:

$$s'(t) = a\cos(\pi(\Xi-\eta)q+\varphi)\times \\ \times\begin{pmatrix}1+2\sum_{k=0}^{+\infty}\cos\pi(\Xi-\eta)(2k+1)\times\\ \times\frac{\sin\pi(2k+1)/2\cos\pi(\Xi+\eta-1/2)(2k+1)}{\pi(2k+1)/2}\end{pmatrix}. \quad (29)$$

Note that replacing the summation index k by $-k-1$ leaves, due to the parity of cosine and the oddness of sine functions, the terms of sum unchanged, while the summation itself is performed from $k = -\infty$ to $k = -1$, which allows, after renaming k back to n, to represent (29), unified for $\pm n$, in the following form

$$s'(t) = a\cos(\pi(\xi-\eta)q+\varphi)\times \times\left(1+\sum_{n=-\infty}^{+\infty}\cos\pi(\xi-\eta)(2n+1)\times \times\frac{\sin\pi(2n+1)/2\cos\pi(\xi+\eta-1/2)(2n+1)}{\pi(2n+1)/2}\right). \quad (30)$$

Further trigonometric transformations (details of which are omitted) turn (30) into the binomial expression (18), this time with a "mother" component

$$s_{\xi,\eta}(\varphi) = \frac{a}{2}\cos(\pi(\xi-\eta)q+\varphi)\times \times\left(1+\cos 2\pi\xi\sum_{n=-\infty}^{+\infty}\frac{\sin 2\pi(2\xi)n}{\pi(n+1/2)}+\sin 2\pi\xi\sum_{n=-\infty}^{+\infty}\frac{\cos 2\pi(2\xi)n}{\pi(n+1/2)}\right). \quad (31)$$

The series in (31) are summed up by formulas (12) where $q-q_0$ is formally replaced by 1/2, and ξ – by 2ξ:

$$s_{\xi,\eta}(\varphi) = (a/2)\cos(\pi(\xi-\eta)q+\varphi)\times \times(1+(\cos 2\pi\xi\cos\pi\{2\xi\})\operatorname{sgn}\{2\xi\}+\sin 2\pi\xi\sin\pi\{2\xi\}). \quad (32)$$

Again, we consider two cases of the sign of 2ξ fractional part in (32) – the main and special ones:

$$s_{\xi,\eta}(\varphi) = \begin{cases} a\dfrac{1+(-1)^{[2\xi]}}{2}\cos(\pi(\xi-\eta)q+\varphi), & \operatorname{sgn}\{2\xi\}=1 \\ (a/2)\cos(\pi(\xi-\eta)q+\varphi), & \operatorname{sgn}\{2\xi\}=0 \end{cases}. \quad (33)$$

Finally, we introduce a fictitious, multiple of 2π, term $-2\pi[\xi]m$ under the cosine, where, as agreed, $m = q-q_0$, and thereby reduce formula (33) to formula (22).

Thus, conversion of the nodal points t_m by interpolation formula (8), and, therefore, (3), is just a special case of signal recovery at all instants of time t.

V. Analysis of the Transformed Signal

The converted signal (18) for all the values of time t, canonical variables ξ, η and signal parameters a, φ, $f = pF/2$ where $p=\xi-\eta$ from formulas (7), looks like a biharmonic signal – the sum of ξ- and η- harmonic components (24)-(25):

$$\begin{aligned} &s'(t) = a_\xi\cos(2\pi(f-[\xi]F)t+\varphi+2\pi[\xi]Ft_0)+ \\ &+a_\eta\cos(2\pi(f+[\eta]F)t+\varphi-2\pi[\eta]Ft_0); \\ &a_\xi = a\frac{1+(-1)^{[2\xi]}\operatorname{sgn}\{2\xi\}}{2},\quad a_\eta = a\frac{1+(-1)^{[2\eta]}\operatorname{sgn}\{2\eta\}}{2}. \end{aligned} \quad (34)$$

The amplitudes of components a_ξ, a_η, depending on the parity of integer numbers $[2\xi]$, $[2\eta]$ and equality to zero or unity of the signs of non-negative numbers $\{2\xi\}$, $\{2\eta\}$, can take one of three values a, $a/2$, 0 each. In the case $[\xi]=-[\eta]=I$, when the cosinusoids coincide, both components are shifted in frequency by $-IF$, in phase – by $2\pi IFt_0$, thus summing into a harmonic signal with amplitude $a_\xi+a_\eta$, which may be equal to a (then at $I=0$ we have the identical transform of signal), $a/2$ (double attenuation of biased (for $I\neq 0$) or unbiased (for $I=0$) signal, $3a/2$ (one-and-a-half-times signal amplification), $2a$ (double gain) and 0 (signal disappearance). These phenomena, which also occur for differing-in-frequency components $I'=[\xi]$, $I''=-[\eta]$, $I'\neq I''$, including the aliasing effect, are shown in Fig. 1 – each pictured in its own color – with their analysis in the following sections *A-E* of current chapter V. The rotation of (ξ, η) plane by 45° clockwise provides, by inversing the formulas (7): $p=\xi-\eta$, $G=\xi+\eta-1$, the display of transform results on (p, G) plane.

A. Conditions for the Theorem Validity

Directly from (34) we can see that condition (4) of signal (1) transform (3) identity will be either zeroing out one of the harmonic components while maintaining the other with amplitude a or two-fold weakening of both components, without frequency and phase offset in each of these cases:

$$\begin{cases} a_\xi = a,\ [\xi]=0,\ a_\eta=0 \\ a_\eta = a,\ [\eta]=0,\ a_\xi=0 \\ a_\xi = a_\eta = a/2,\ [\xi]=[\eta]=0 \end{cases}. \quad (35)$$

The amplitude of ξ-component in the top line of collections of equalities (35) will be equal to a if (-1) in (34) is raised to an even power: $[2\xi]=2k$; moreover, $\{2\xi\}\neq 0$ and there is no bias, i.e., $[\xi]=0$. These conditions are formulated as a system of inequalities: $2k\le 2\xi<2k+1$, $0\le\xi<1$, $2\xi\neq 2k$, the solution of which is the inequality $0<\xi<1/2$. For the η-component to be zero, (-1) is raised to an odd power: $[2\eta]=2l-1$, while $\{2\eta\}\neq 0$ regardless of the presence or absence of bias, this formulated as a system $2l-1\le 2\eta<2l$, $2\eta\neq 2l-1$, the solution of which is the inequality $l-1/2<\eta<l$. By swapping ξ and η, as well as k and l, we establish that the second line (35) is equivalent to similar system of inequalities $0<\eta<1/2$, $k-1/2<\xi<k$. In the third line (35), the condition of the absence of bias $[\xi]=[\eta]=0$ gives the system of inequalities $0\le\xi<1$, $0\le\eta<1$, or, $0\le 2\xi<2$, $0\le 2\eta<2$, whence we have $[2\xi]=0$ or 1, $[2\eta]=0$ or 1, which, together with the condition of equality of both amplitudes $a/2$: $\{2\xi\}=\{2\eta\}=0$ meaning $2\xi=[2\xi]$, $2\eta=[2\eta]$, defines four isolated points on the (ξ, η) plane: (0, 0), (0, 1/2), (1/2, 0), (1/2, 1/2).

Thus, the domains of theorem's validity are described by the collection of inequalities and equalities

$$\begin{bmatrix} 0<\xi<1/2,\ l-1/2<\eta<l;\ 0<\eta<1/2,\ k-1/2<\xi<k \\ \xi=0,\ \eta=0;\ \xi=0,\ \eta=1/2;\ \xi=1/2,\ \eta=0;\ \xi=1/2,\ \eta=1/2 \end{bmatrix} \quad (36)$$

where the comma and semicolon mean logical "and" and "or", respectively – with priority '∧' over '∨'.

The upper line in collection (36) defines in Fig. 1 the green squares without external borders, the lower one – the specified four points of bright green color. The integer variables k and l number the squares located along the horizontal and vertical axes of the (ξ, η) plane, respectively. The segment of bright green color $0\le p<1$ at $G=0$ (in the form of an arrow including the starting point and excluding the final) corresponds to the classical formulation of theorem [1], which states, according to our notation $f=pF/2$, that the signal (5), the spectrum of which is located in frequency band $0\le f<F/2$, is restored from

its readings obtained with sampling frequency F by the interpolation formula (3) with the frequency band index $G=0$. It can be seen that for other integer values of G the frequency bands where the theorem holds are also represented by diagonals of green squares parallel to the p axis: $GF/2<f<(G+1)F/2$ (hence the name of parameter G – index of the frequency band). The frequency range $F_{low}<|f|<F_{up}$ where the theorem holds in the general case of real G as will be found later (see (44) and section VI.*B*), is determined by formulas

$$\begin{cases} F_{low} = ([|G+1/2|] - |\{G\} - 1/2| + 1/2)F/2 \\ F_{up} = ([|G+1/2|] + |\{G\} - 1/2| + 1/2)F/2 \end{cases}. \quad (37)$$

The range is of maximum width for integer index G: $F_{up}-F_{low}=F/2$, and degenerates for half-integer: $F_{up}-F_{low}=0$.

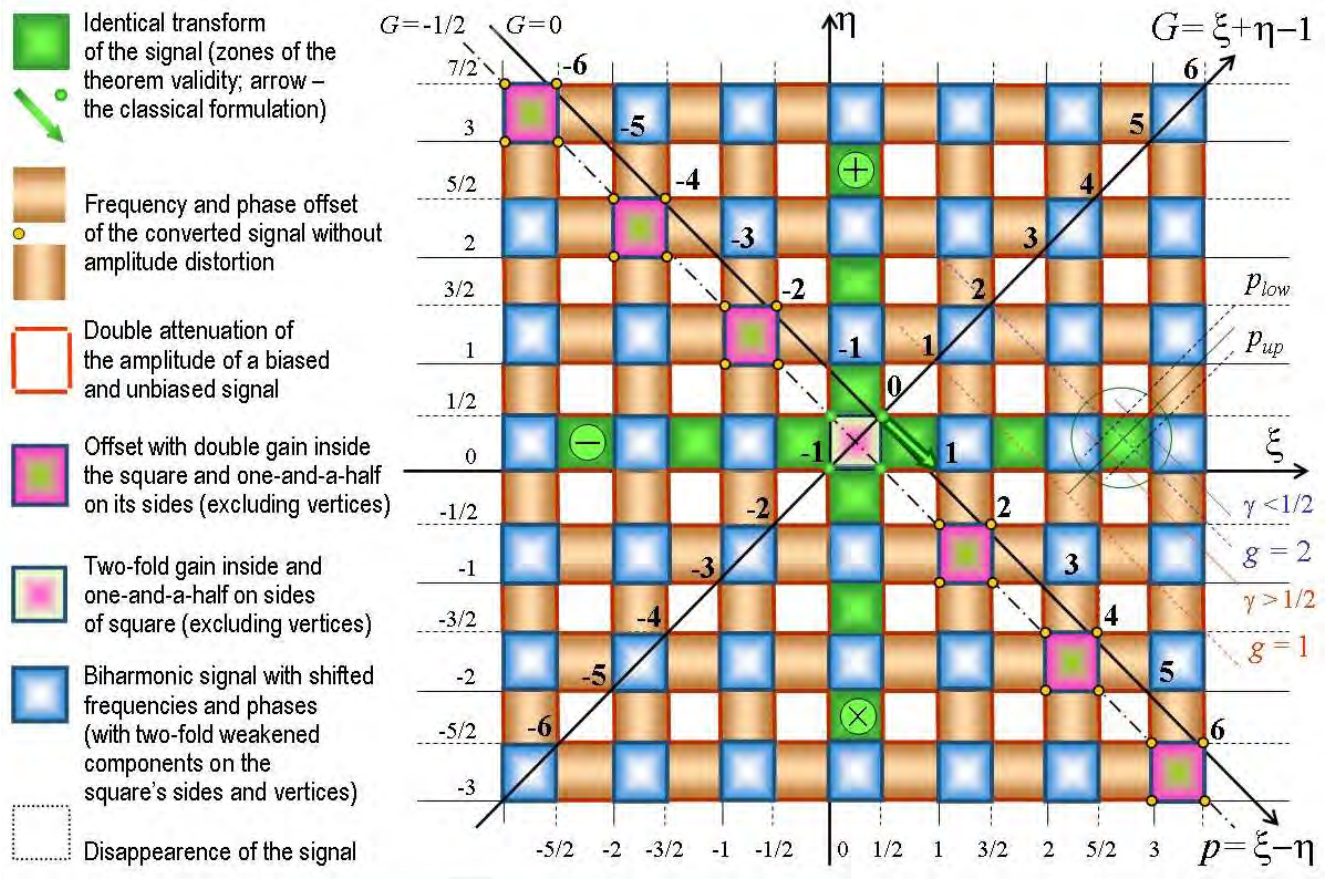


Fig. 1. Geometric representation of the results of conversion of infinite duration harmonic signal by interpolation formula of sampling theorem in time domain

B. Double and One-and-a-half-times Amplification and Double Signal Attenuation

The converted signal (34) can have the same frequency f and phase φ as the original (1), but amplitude that differs from a – in compliance with the phenomena of theorem violation named in the heading of this section:

$$\begin{cases} ([\xi]=0 \wedge a_\xi = a/2 \wedge a_\eta = 0) \vee ([\eta]=0 \wedge a_\eta = a/2 \wedge a_\xi = 0) \\ [\xi]=[\eta]=0 \\ (a_\xi = a/2 \wedge a_\eta = a) \vee (a_\xi = a \wedge a_\eta = a/2) \vee (a_\xi = a \wedge a_\eta = a) \end{cases}. \quad (38)$$

The first line of collection (38) defines the conditions for double attenuation: the amplitude of one of the components is $a/2$ (in absence of frequency and phase offset), and the other component is zero (regardless of which offset). The system of the following two lines permits the existence of both components with amplitudes a and/or $a/2$, but without frequency and phase offset for both components. The equality of the amplitudes a_ξ and a_η to zero or a we analyzed above (in the section V.*A*). The equalities $a_\xi=a/2$, $a_\eta=a/2$ hold for $\operatorname{sgn}\{2\xi\}=0$, $\operatorname{sgn}\{2\eta\}=0$ – when ξ,η are integers or half-integers. The formulation of these conditions in the language of inequalities, their equivalent transformations (the details of which are omitted), lead (38) to the following collection of relationships with respect to ξ and η:

$$\begin{cases} (\xi = 0 \vee \xi = 1/2) \wedge (0 < \eta < 1/2 \vee l - 1/2 < \eta < l) \\ (\eta = 0 \vee \eta = 1/2) \wedge (0 < \xi < 1/2 \vee k - 1/2 < \xi < k) \\ 0 < \xi < 1/2 \wedge 0 < \eta < 1/2 \end{cases}. \quad (39)$$

The first line (39) describes in Fig. 1 two straight vertical segments of 1/2 length (with excluded ends) of blue color, where one-and-a-half-times signal amplification takes place, as well as a sequence of the same type equidistant segments of red color (numbered by $l=0,\pm1,\pm2,\ldots$) where double signal attenuation occurs. If in the description just made we swap the symbols ξ and η, as well as l and k, then, in accordance with

second line (39) exactly the same segments, only horizontal ones, are illustrated in Fig. 1. Finally, the third line describes the central square inside which the signal is amplified twice.

C. Signal's Frequency and Phase Offset

The converted signal is shifted in frequency and phase compared to the original under the following conditions:

$$\begin{cases} [\xi]\neq 0 \wedge ([2\xi]\bmod 2 = 0 \vee \{2\xi\} = 0) \wedge ([2\eta]\bmod 2 = 1 \wedge \{2\eta\} \neq 0) \\ [\eta]\neq 0 \wedge ([2\eta]\bmod 2 = 0 \vee \{2\eta\} = 0) \wedge ([2\xi]\bmod 2 = 1 \wedge \{2\xi\} \neq 0) \\ ([2\xi]\bmod 2 = 0 \vee \{2\xi\} = 0) \wedge ([2\eta]\bmod 2 = 0 \vee \{2\eta\} = 0) \wedge ([\xi] = -[\eta] \neq 0) \end{cases}. \quad (40)$$

The first and second lines in collection (40) indicate that the signal is pure harmonic: it consists of either ξ-component ($[\xi]\neq 0$, $a_\xi \neq 0$, $a_\eta = 0$), or η-component ($[\eta]\neq 0$, $a_\eta \neq 0$, $a_\xi = 0$). The third line allows for the presence of two components, provided that both are offset by the same non-zero frequency and phase ($a_\xi \neq 0$, $a_\eta \neq 0$, $[\xi] = -[\eta] \neq 0$). The offset can be accompanied by the signal amplification and attenuation phenomena described in previous section.

After writing down the operations of taking integer and fractional parts in (40) in the language of inequalities and performing equivalent transformations on them (the details of which are omitted), we arrive at a system of inequalities with integers k and l independent of each other:

$$\begin{cases} \begin{bmatrix} k \le \xi \le k+1/2 \wedge l-1/2 < \eta < l \\ k \le \eta \le k+1/2 \wedge l-1/2 < \xi < l \\ k \le \xi \le k+1/2 \wedge -k \le \eta \le -k+1/2 \end{bmatrix} \\ k = \pm 1, \pm 2, \ldots; \quad l = 0, \pm 1, \pm 2, \ldots \end{cases}. \quad (41)$$

The net offset, when amplitude a is preserved (with inequality '<' in (41) strict for k and l instead of non-strict '≤'), is shown in Fig. 1 by yellow-brown squares. Attenuation of the shifted signal is represented by red-colored sides of these squares. Double amplification takes place inside pink squares located diagonally along the p axis, and one-and-a-half-times amplification occurs on their blue sides, while at the vertices of squares – at isolated yellow points – the amplitude is not distorted.

D. Imposition of Two Signal Components with Different Frequencies

The conditions $a_\xi \neq 0$ and $a_\eta \neq 0$ for $[\xi] \neq -[\eta]$ mean that the converted signal (34) consists of two differing in their frequency components. Such a splitting of source signal (3), as shown in [8], creates the well-known frequencies overlay effect at the edges of frequency range $[G]\,F/2 < f < [G+1]\,F/2$ with a non-integer index $G > 0$ (a complete description of this phenomenon for all values of G and f is provided by the system of relations (45)). The overlay conditions formulated here in the language of inequalities look like this: $k \le \xi \le k+1/2$, $l \le \eta \le l+1/2$, $k \neq -l$ where $k, l = 0, \pm 1, \pm 2, \ldots$. The overlay zones in Fig. 1 are represented by blue squares with the amplitudes of components $a_\xi = a_\eta = a$ inside, $a_\xi = 2a_\eta = a$ and $a_\eta = 2a_\xi = a$ on horizontal and vertical sides, respectively, $a_\xi = a_\eta = a/2$ at vertices.

E. Complete Disappearance of the Signal

This paradoxical phenomenon, first described in the literature by the author [8], is obvious from formula (34): $a_\xi = a_\eta = 0$. This double equality is reduced (as shown above for cases of amplitude zeroing) to the system of inequalities $k - 1/2 < \xi < k$, $l - 1/2 < \eta < l$ where k and l are arbitrary integers independent of each other. Each pair (k, l) defines a white square with excluded sides and vertices in Fig. 1.

VI. Comparison with the Theorem for a Signal of Finite Duration

From both mathematical and physical points of view, our theorem can be considered as a special case of the time domain sampling theorem for a finite signal – when its duration T tends to infinity while the width F of its spectral function is staying constant. In the formulation for a finite number of samples $N=FT$ transformed by the formula

$$s'(t) = \sum_{n=0}^{N-1} s(t_n) \frac{\sin \pi (G+1) F(t-t_n) - \sin \pi G F(t-t_n)}{N \sin \pi F(t-t_n)/N} \quad (42)$$

it was proved in the works of the author [7]-[8]. It should be noted that the method developed there for calculating the finite sum (42) radically differs from summing up an infinite series (3) by ready-made formulas (12). For a finite number of samples, it is principally impossible to derive an explicit two-term analytical formula of type (34), but one can only represent $s'(t)$ as an expansion into infinite series of harmonics – replicas of the original signal (1), shifted in frequency and phase, no more than two of which have a non-zero amplitude. Conditions for the theorem validity, the phenomena of two signals imposition, displacement of the monochromatic and disappearance of the entire signal are expressed in the form of inequalities, which include an indefinite integer parameter I – the order number of converted signal's branch. Only now it became clear that the numbers of branches appearing in these inequalities (see (43), (45), (46)) are none other than $I = I' = [\xi]$, $I'' = -[\eta]$ in subcosine expressions of formula (34). Without the derivation of this explicit formula, of course, we could not speak about the canonical variables ξ and η, which represent the signal conversion results in the form of a "chessboard" in Fig. 1, the colored cells of which show an amazing symmetry and equal status of parameters p and G – the current signal frequency and the index of range where this frequency is located.

A. Separate Branches of the Converted Signal

We present the main result of the finite version of theorem: an inequality that connects p and G when the I-th branch of signal exists in its pure form (without imposing another component on it):

$$\begin{cases} s'(t) = a\cos(2\pi(f - IF)t + \varphi + 2\pi IFt_0) \\ \| f/F - I \| < (([|G+1/2|]+1/2)/2 | < (| \{G\} - 1/2| + 1/N)/2 \end{cases}. \quad (43)$$

Assuming $I = 0$, for $N \to \infty$ we obtain the inequality

$$\| p | < [| G + 1/2 |] - 1/2 | < | \{G\} - 1/2 | \quad (44)$$

defining the limits of applicability of our theorem announced by formulas (37) and arranged in the first line (36) as a collection of inequalities describing the green squares numbered by integers k and l in Fig. 1. In order to verify both theorem versions, we will try to deduce the very non-trivial relationship (44) from the geometric constructions on Fig. 1.

B. Derivation of Analytical Expression of Theorem Validity from a Geometric Illustration

Consider a green square surrounded by a circle pierced by two level lines $G = g + \gamma$ parallel to its diagonal, where $g = [G]$ is the integer part of band index, $\gamma = \{G\}$ – its fractional part. These lines represent two characteristic cases for γ: $\gamma < 1/2$ (blue color of the inscription) and $\gamma > 1/2$ (red). For this selected square, in the first case $g = 2$, in the second – $g = 1$. The theorem is valid inside the square on two parallel segments $p_{low} < p < p_{up}$, the boundaries of which $p_{low} = 2F_{low}/F$ and $p_{up} = 2F_{up}/F$ are intersection points of the line $G = 2 + \gamma$ on plane (p, G) with lines $\eta = 1/2$, $\xi = 3 = k$ on plane (ξ, η) for the case $\gamma < 1/2$, and the line $G = 1 + \gamma$ with lines $\xi = 5/2 = k - 1/2$, $\eta = 0$ for the case $\gamma > 1/2$. In the first case, from the equation $\eta = (G - p_{low} + 1)/2 = 1/2$, given by second formula (7), we find $p_{low} = G = g + \gamma = 2 + \gamma$. The upper cut-off frequency p_{up} is determined from the similarity of isosceles right-angled triangles with bases $p_{up} - p_{low}$ and 1 (diagonal of the square) and heights $1/2 - \gamma$ and $1/2$ (the half of other diagonal). From the equality of ratios $(p_{up} - p_{low})/1 = (1/2 - \gamma)/(1/2)$ for lengths of specified bases and heights, we find $p_{up} = p_{low} + 1 - 2\gamma = g + 1 - \gamma$. In the second case, the same formula (7) sets equation $\eta = (G - p_{up} + 1)/2 = 0$, whence we determine the upper cut-off frequency $p_{up} = G + 1 = g + 1 + \gamma$. The lower cut-off frequency is found from the similarity of triangles of the above type, but located below the diagonal of that square: from the proportion $(p_{up} - p_{low})/1 = (\gamma - 1/2)/(1/2)$ we get $p_{low} = p_{up} - (2\gamma - 1) = g + 2 - \gamma$. Further, we notice that a square marked with '+' sign in the upper half-plane is symmetrical to the square just used with respect to G axis, i.e., the theorem is also valid in frequency band $-p_{up} < p < -p_{low}$ with the same g and γ, which allows us to write the frequency domain set for both squares in the form of inequality $p_{low} < |p| < p_{up}$. The remaining two squares marked with signs '×' and '–' are symmetrical to those just considered with respect to the straight dash-dotted line $G = -1/2$ (the reason for this symmetry lies in the fact that both transforms (3) and (42) are invariant when replacing $G + 1/2$ with $-G - 1/2$, or, G by $-G - 1$). But then in the range generalized over all four squares, instead of $G + 1/2$, for the same γ, everywhere should be written $|G + 1/2|$. In the first case, we proceed from the range $g + \gamma < |p| < g + 1 - \gamma$, in the second – from $g + 2 - \gamma < |p| < g + 1 + \gamma$. Subtracting g from the first inequality and $g + 1$ from the second, we obtain a collection of two systems of inequalities describing both cases for γ: $\gamma < |p| - g - 0 < 1 - \gamma$, $\gamma < 1/2$; $1 - \gamma < |p| - g - 1 < \gamma$; $\gamma > 1/2$. We generalize both systems using the identities $[\gamma + 1/2] = 0$, $1/2 - |\gamma - 1/2| = \gamma$ valid for $\gamma < 1/2$ and the identities $[\gamma + 1/2] = 1$, $1/2 - |\gamma - 1/2| = 1 - \gamma$ true for $\gamma > 1/2$: $1/2 - |\gamma - 1/2| < |p| - g - [\gamma + 1/2] < 1 - (1/2 - |\gamma - 1/2|)$. The resulting inequality is reduced, by equivalent transformations, to inequality $||p| - [G + 1/2] - 1/2| < |\{G\} - 1/2|$ where it remains to enclose the expression in square brackets into modular brackets in order to arrive at the inequality (44), to which should be added four isolated points ($p = 0$, $G = -1/2 \pm 1/2$), ($p = \pm 1/2$, $G = -1/2$) of a bright green color at the vertices of central square. Note that the inequality (43) for $I = 0$ does not need such an addition: it is satisfied for all the specified four points, though none of them is allowed by the computability conditions for transform (42) when N is even.

C. Imposition of Branches and Signal Disappearance

The geometric method shown in the previous section is hardly applicable for verification of inequalities obtained in [8] by a detailed analysis of relationships describing the conditions of two branches overlay

$$\begin{cases} s'(t) = a\cos(2\pi(f - I'F)t + \varphi + 2\pi I'Ft_0) + \\ + a\cos(2\pi(f - I''F)t + \varphi + 2\pi I''Ft_0) \\ |f/F - (I' + I'')/2| \le (1/2 - |\{G\} - 1/2| - 1/N)/2 \\ |I' - I''| = |[G] + 1| \end{cases} \quad (45)$$

and of complete signal disappearance

$$\begin{cases} s'(t) = 0 \\ ||f/F - I - 1/2| - |[G] + 1|/2| \le (1/2 - |\{G\} - 1/2| - 1/N)/2 \end{cases}. \quad (46)$$

But the transition to canonical variables (7) when $N \to \infty$, taking into account the equalities $I' = [\xi]$, $I'' = -[\eta]$, $2f/F = p = \xi - \eta$ set above, represents the system consisting of second and third relationships (45) in an elegant form:

$$\begin{cases} |\{\xi\} - \{\eta\}| + |\{\xi + \eta\} - 1/2| \le 1/2 \\ |[\xi] + [\eta]| = |[\xi + \eta]| \end{cases}. \quad (47)$$

It turns out that the inequality in the first line (47) colors the plane (ξ, η) into a "chessboard" with fields $1/2 \times 1/2$ – black where it is true, and white where it is false. Equality in the second line repaints half of these fields in white (where it is incompatible with inequality). The final picture is presented in Fig. 1 where the role of remaining black fields is played by squares of blue color (overlapping zones of two different frequency harmonics) and zones of "autoimposition" – squares of pink color (amplitude amplification with frequency and phase offset), and the central square (amplification without offset).

In view of the importance of system (47), the application of which is not limited to the sampling theorem, but represents a universal technique for geometric illustration of analytical expressions, we give its proof starting with inequality in the upper line. First of all, note that this inequality is periodic in ξ and η with period of 1: it does not change when replacing ξ by $\xi + k$ and η by $\eta + l$ where k and l are arbitrary integers. Therefore, it is sufficient to examine it in the unit (base) square E ($0 \le \xi < 1$, $0 \le \eta < 1$) the image of which can then be repeated on the (ξ, η) plane an infinite number of times. We consider the first half of this square – triangle D with vertices (0,0), (0,1), (1,0), which consists of a smaller square C ($0 \le \xi \le 1/2$, $0 \le \eta \le 1/2$) and two smaller triangles A and B adjacent to it on the right and above. The set of points $(\xi,\eta) \in D$, as we know from analytical geometry, is described by inequality $\xi + \eta \le 1$, which allows us to write inequality (47) without curly brackets: $|\xi - \eta| + |\xi + \eta - 1/2| \le 1/2$. By squaring both non-negative parts of it and performing equivalent transformations, we get inequality $(\xi - 1/2)\xi + (\eta - 1/2)\eta \le 0$ which is true in C due to the non-positivity of both its left-hand terms. As for A and B, the sign of this left part is not defined in them, but the sign of the submodule expression is defined clearly: $\xi + \eta - 1/2 > 0$. This allows us to write our inequality as $|\xi - \eta| + \xi + \eta \le 1$, or $\xi \le 1/2$ in A and $\eta \le 1/2$ in B, which contradicts the definition domains $1/2 < \xi < 1$ and $1/2 < \eta < 1$ of variables ξ and η in A and B. Thus, inequality

(47) is true in C and false in A and B. The second part of the base square E is triangle D' with vertices (1,1), (0,1), (1,0) which is symmetric to D with respect to the same diagonal of E, so that each point $(\xi,\eta)\in D'$ is mirrored at the point $(\xi',\eta')\in D$ where $\xi'=1-\xi$, $\eta'=1-\eta$. Moreover, inequality $\xi+\eta>1$ describing the domain D' turns into $\xi'+\eta'<1$ for D. In addition, inequality (47), in view of $\{\xi+\eta\}=\xi+\eta-1=1-\xi'-\eta'$, takes the form $|\xi'-\eta'|+|\xi'+\eta'-1/2|\le 1/2$, i.e., remains invariant in D. This means, by virtue of the indicated symmetry, that it is true in the square $C'(1/2\le\xi<1,\ 1/2\le\eta<1)$ and false in the triangles A' and B' that are mirror-wise opposite to A and B with respect to the diagonal of square E, thus making-up the chessboard coloring of plane (ξ,η) by means of inequality (47). We now proceed to the equality in the second line (47). It is true at all internal points of triangle D, on its cathets, at the right angle's vertex – where schematically it looks like $|0+0|=|0|$, as well as in the vertices of acute angles: $|0+1|=|1|$; on hypotenuse, the equality is false: $|0+0|=|1|$. Triangle D' has a different picture: the equality is true on the cathets and at the apex of the right angle: $|1+0|=|1|$, $|1+1|=|2|$, and is false at all internal points: $|0+0|=|1|$. All the above statements retain their logical values under the periodic continuation $\xi\to\xi+k$, $\eta\to\eta+l$: $|0+0|=|0|\to|k+l|=|k+l|$, $|0+0|=|1|\to|k+l|=|k+l+1|$ and so on. It turns out that equality (47) covers the plane (ξ,η) with a "tile" of triangular shape in two colors – black D and white D', so it acquires an independent value as one more manner of analytical and geometric description of two-dimensional periodic structures. As a result, black squares with the C' prototype are excluded from the chessboard, as required for illustrating the imposition and auto-imposition zones according to formula (45) – taking into account the inclusion into the system (47) of missing blue squares vertices (with the $\xi=\eta=1/2$ prototype – in the middle of triangle's D hypotenuse – where the system is inconsistent) and excluding the vertices of the central and pink squares.

It can be seen from Fig. 1 that the signal disappearance zones are arranged in the same "square-nested" order as the frequency overlay zones, but they are shifted with respect to the latter by 1/2 for ξ and 1/2 for η. Their analytical description is given by the same system (47) where the inequality in the first line has a strict sign '<' (to exclude the contours of chessboard black fields), and the equality in the second line is replaced by its negation: $|[\xi]+[\eta]|\ne|[\xi+\eta]|$ – in order to invert the colors of black and white fields having the prototypes C and C', respectively.

D. Conditions for the theorem validity in canonical variables

Based on the results obtained in the previous sections of this chapter, we can formulate, proceeding just from the geometric illustration in Fig. 1, the conditions for the validity of our theorem in the notations of canonical variables only:

$$\left[\begin{array}{l}\begin{cases}|\{\xi\}-\{\eta\}|+|\{\xi+\eta\}-1/2|\ge 1/2\\ \{2\xi\}\{2\eta\}\ne 0\\ [2\xi][2\eta]=0\end{cases}\\ \begin{cases}\{2\xi\}+\{2\eta\}=0\\ |[\xi]|+|[\eta]|=0\end{cases}\end{array}\right. . \qquad (48)$$

Inequality in the upper system of collection (48), being a non-strict negation of inequality in system (47), also paints the plane (ξ,η) in chessboard order but inverted in color: black fields represent zones of signal monochromaticity (43), white fields represent the imposition and auto-imposition zones (45), as well as zones of signal disappearance (46). The second inequality of this system excludes the boundaries (sides and vertices) of all board fields – where the signal amplitude is distorted.

To separate the zones of theorem validity from those zones of monochromaticity where the signal is shifted, an equality is used in the same upper system of collection (48), which states that there is no bias for either vertically or horizontally arranged green squares on the plane.

The lower system in collection (48) describes four isolated points of bright green color in Fig. 1. The first double equality of this system means that ξ and η are integers and half-integers, and, thus, indicates the vertices of all the squares of plane (ξ,η) without exception. The second double equality selects from this discrete set of points the required vertices of the central square.

E. Absence of Half Amplitudes on the Boundary Lines in the Finite Version of the Theorem

Interestingly, when converting a finite duration signal, the amplitudes of components a_ξ and a_η turn out to be equal only to a and/or 0, but never to $a/2$. This means that double attenuation and one-and-a-half-times amplification do not occur in the finite version of theorem. Let us present this remarkable fact as a special case of formula (34) for infinite version, assuming that a_ξ and a_η are obtained under restrictions on the finite. It was established in [7] that the finite sum (42) can be expanded into the sum of harmonic components not for any real values of G and p, but only for rational $G=g+l/N$; $l=0,1,\ldots,N-1$ and $p=(2m+j)/N$; $m=0,\pm1,\pm2,\ldots$; $j=0$ or 1. In the case of odd $N=2M+1$; $M=0,1,2,\ldots$ the index G can only be equal to 0 or -1 (the "classical" frequency range), and j – to zero. Hence, 2ξ, 2η in the formulas (34) for amplitudes a_ξ, a_η are $G\pm p+1=\pm 2m/(2M+1)$, and may not be integers providing their fractional parts to be zero (for equality the amplitude to $a/2$), since both represent a fraction, the numerator of which is even, the denominator – an odd number. In the case of even $N=2M$ the indicator j is equal to $1-l \bmod 2$ that according to the first formula (7) for $G\ge -1/2$ gives $2\xi=g+l/N+(2m+1-l+2[l/2])/N+1=(2L+1)/(2M)$ where $L=(g+1)M+m+[l/2]$ is an integer number. Exactly the same expression is obtained from the second formula (7) for 2η, but with integer $L=(g+1)M-m-1+l-[l/2]$. The case $G<-1/2$ is satisfied for 2ξ and 2η for integer $L=-gM+m-l+[l/2]$ and integer $L=-gM-m-1-[l/2]$, respectively. Again, 2ξ and 2η are fractions with different parity of numerator and denominator, and therefore the amplitudes a_ξ and a_η in (34) can only be equal to a or 0.

F. The Full Illustration of the Finite Version of the Theorem

Fig. 2 shows the results of converting the finite duration harmonic signal (42) in the form of discrete points on (p, G) plane, the colors of which reflect all the computable phenomena: theorem validity (44), signal frequency and phase offset (43), imposition and autoimposition of frequencies (45), signal disappearance (46). The picture was obtained by numerical modeling of these formulas, and, actually, verifies the infinite version of theorem, illustrated in Fig. 1.

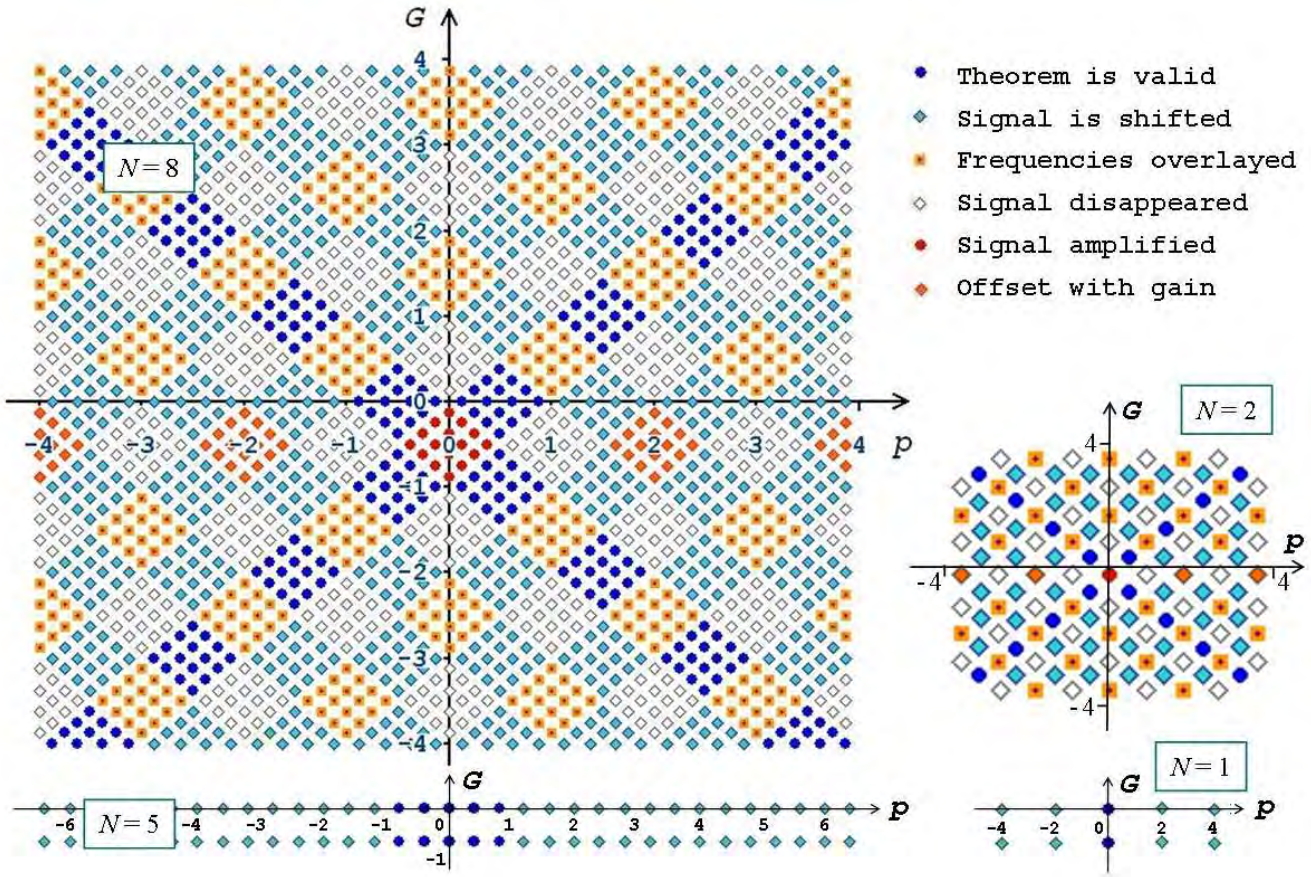


Fig. 2. Geometric representation of the results of conversion of finite duration harmonic signal by interpolation formula of sampling theorem in time domain

VII. CONCLUSION

A time-domain sampling theorem has been formulated and proved for a process of infinite duration occurring in a limited frequency band with an arbitrary index *G*.

The zones of the phenomena of validity and violation of the theorem – identical signal transform (the theorem is true), frequency overlay, double attenuation, doubling or one-and-a-half-times amplification with or without the frequency and phase offset, disappearance of the signal fill the entire plane of parameters (p,G), the rotation of the axes of which by 45° represents the same results on the plane of canonical variables (ξ, η) in a "staggered" order with greater clarity, both their analytical expressions and geometric locations.

A comparison is made of the conversion results of infinite and finite duration signals, which showed that both versions of the theorem – infinite and finite – are special cases of each other.

The representation of signal dimensionless frequency and its variation range index in the form of peer canonical variables extends the theoretical significance of sampling theorem beyond its scope of applications for radio and communication engineering, data compression, interpolation, etc., and makes it possible to formulate fundamental physical principles in the language of Fourier analysis statements.

REFERENCES

[1] V.A. Kotelnikov, "On the carrying capacity of "ether" and wire in telecommunications," Proceedings of 1st all-union congress on the technical reconstruction of communication and the development of low-current industry, Izd. Red. Upr. Svyazi RKKA, Moscow, 1933 (in Russian).

[2] D.A. Linden, "A discussion of sampling theorems," Proc. IRE, vol. 47, July 1959, pp. 1219-1226.

[3] Liu Jianhua, Zhou Xiyuan, and Peng Yingning, "Spectral arrangement and other topics in first-order bandpass sampling theory," in IEEE Trans. Signal Process., vol. 49, no. 6, June 2001, pp. 1260-1263.

[4] V.V. Petrov, A.S. Uskov, Information theory of synthesis of optimal control and management systems (continuous systems), Moscow: "Energia", 1975, 232 p. (in Russian).

[5] A.J. Jerri, "The Shannon sampling theorem – its various extensions and applications: a tutorial review," Proc. IEEE, vol. 65, no. 11, November 1977, pp. 1565-1596.

[6] G.S. Khanyan, "Generalization of sampling theorem to the case of non-integer index of the band", International scientific conference on the 100th anniversary of V.A. Kotelnikov: Moscow, 21-23 October 2008, Abstracts, Publishing House of MPEI, Moscow, 2008, pp. 35-37 (in Russian).

[7] G.S. Khanyan, "Sampling theorem for finite duration signal with a non-obligatory zero index of the frequency band," Izvestiya Yuzhn. Feder. Universiteta. Tekhnicheskie nauki, 2013, no. 2 (139), pp. 20-25 (in Russian).

[8] G.S. Khanyan. Features of limited duration harmonic signal transform by sampling theorem // Problems of advanced micro- and nanoelectronic systems development, 2017, Part I, Moscow, IPPM RAS, pp. 54-60 (translated from Russian).

[9] A.P. Prudnikov, Yu.A. Brychkov, O.I. Marichev, Integrals and series. Elementary functions. Moscow: "Nauka", 1981, 800 p. (in Russian).

Structure of the Transfer Function Numerator Coefficients as One of the Factors of the Structural Precision of IIR Digital Filters

Vladislav Lesnikov
Department of Radio Electronic Systems
Vyatka State University
Kirov, Russia
Vladislav.Lesnikov.Ru@IEEE.org

Tatiana Naumovich
Department of Radio Electronic Systems
Vyatka State University
Kirov, Russia
NTV_new@mail.ru

Alexander Chastikov
Department of Radio Electronic Systems
Vyatka State University
Kirov, Russia
AlChast@mail.ru

***Abstract*— This work was carried out during the implementation of a project dedicated to the synthesis of recursive digital filters with a finite word length. The approach used in this project considers the finite word length already when calculating the zeros and poles. The calculated zeros and poles are not distorted by structural synthesis. The calculations are based on what the zeros and poles of a digital filter with finite length coefficients are algebraic numbers of the corresponding degree. Structural synthesis involves generating a topological matrix describing the structure. The purpose of the generation is to provide a given degree of algebraic numbers that are zeros and poles. The previously calculated values of zeros and poles are used to calculate the exact values of the coefficients of the generated structure. It is proposed to characterize various structures with structural precision that does not depend on the concrete values of the coefficients. This article discusses the issues of generating structures considering the structural precision of the transfer function numerator.**

Keywords*— *IIR digital filters, finite word length, number theoretic approach, matrix structure description, constituent units of structural accuracy

I. Introduction

The use of digital filters with infinite impulse responce (IIR DF) has a very long history [1]. However, the problem of correct design of IIR DF with stringent requirements of specifications, taking into account the finite word length (FWL), remains still far from the final solution. The top FPGA companies (Xilinx, Intel FPGA) have in their disposal software products such as finite impulse response (FIR) compilers. An attempt is known to develop and use a layered product for IIR DF by Altera Corporation (IIR Compiler). However, it was unsuccessful and support for this system was discontinued in 2003 [2]. The reason for these failures is that traditional design methods [3] – [5] lack the depth of knowledge about recursive digital filtering processes.

The authors of this paper have conducted research that allowed them to acquire in-depth knowledge of the number-theoretic and algebraic nature of IIR DF [6]. The results obtained have been incorporated as a basis for the new paradigm to the design of IIR FWL filters [7], [8].

First, under this approach, the finite word length is taken into consideration for the final calculation of zeros and poles even before the stage of structural synthesis [9], [10]. This means that the structural synthesis step does not distort the previously calculated values of the zeros and poles.

Secondly, structures are generated to ensure the exact implementation of the calculated zeros and poles [11], [12].

To reduce the number of options for generating structures, the proposed approach involves the use of the concept of structural precision. This concept allows to compare different structures in accuracy without specifying filter parameters. One aspect of structural precision is the so-called structure of the transfer function coefficients of a particular filter block diagram. In [13], the structure of the denominator coefficients of the transfer function is considered. This paper explores the structure of the transfer function numerator coefficients.

II. Relationship between the Transfer Function and the Matrix Description of DF Structure

The digital filtering algorithm is based on multiple cyclic repetition of a set of operations describing the calculation of the next sample of the output sequence. The set of these operations are usually depicted graphically in the form of a block diagram of a digital filter.

An adequate mathematical model of the filter block diagram is a weighted directed graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, N vertices of which $v_i \in \mathbf{V}$ ($i = 1, ..., N$) correspond to the nodes of the block diagram, and the weights $e_{ij} \in \mathbf{E}$ of the edges of which correspond to the transmission coefficients t_{ij} from the node with the number j to the node with the number i (it is assumed that the transmission coefficient is calculated for z-transforms of the sample sequences calculated at nodes). The graph $\mathbf{G}$ is described by an adjacency matrix [14], which will be called a topological matrix.

Fig. 1 shows an example of a block diagram of a third-order DF.

The condition for the physical realizability (computability) of a DF with a topological matrix $\mathbf{T}(z)$ is the existence of such a

The reported study was funded by RFBR, project number 18-07-00986 A.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

numbering of the nodes that all elements of the topological matrix, other than 0 and z^{-1}, must be located below the main diagonal [14]. In the block diagrams of such DF, there are no loops without delay blocks.

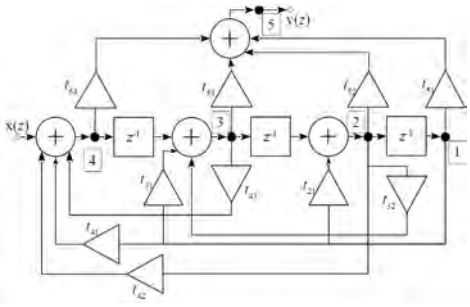


Fig. 1. An example of a DF block diagram (DF order $n = 3$, number of nodes $N = 5$)

The topological matrix of the scheme shown in Fig. 1 is as follows:

$$\mathbf{T}(z) = \begin{bmatrix} 0 & z^{-1} & 0 & 0 & 0 \\ t_{21} & 0 & z^{-1} & 0 & 0 \\ t_{31} & t_{32} & 0 & z^{-1} & 0 \\ t_{41} & t_{42} & t_{43} & 0 & 0 \\ t_{51} & t_{52} & t_{53} & t_{54} & 0 \end{bmatrix}. \quad (1)$$

The operation of filters with an arbitrary topological matrix $\mathbf{T}(z)$ and with arbitrary input and output nodes is described by the matrix equation

$$\mathbf{Y}(z) = \mathbf{T}(z)\mathbf{Y}(z) + \mathbf{I}\,\mathrm{x}(z), \quad (2)$$

where $\mathrm{x}(z)$ is the z-transform of the input sequence, $\mathbf{Y}(z) = [\mathrm{y}_{ij}(z)]$ is the matrix of the z-transforms of the sequences calculated at the i th nodes of the structure, if the input sequence is fed to the j th node, $\mathbf{I}$ is the identity matrix. From equation (2) it is easy to obtain an expression for the matrix of transfer functions

$$\mathbf{H}(z) = \left[\mathrm{H}_{ij}(z)\right] = \frac{1}{\mathrm{x}(z)}\mathbf{Y}(z) = \left(\mathbf{I} - \mathbf{T}(z)\right)^{-1}, \quad (3)$$

where $\mathrm{H}_{ij}(z)$ is the transfer function of the DF, described by the topological matrix $\mathbf{T}(z)$, and i and j are the numbers of the output and input nodes, respectively.

The transfer function of an arbitrary n th order DF is described by the expression

$$\mathrm{H}(z) = \sum_{i=0}^{n} b_i z^{n-i} \Bigg/ \left(z^n + \sum_{i=1}^{n} a_i z^{n-i} \right). \quad (4)$$

Equation (3) makes it possible to establish the relationship between the coefficients of the numerator (b_i) and denominator (a_i) of (4) and the coefficients t_{ij} of the block diagram (elements of the topological matrix $\mathbf{T}(z)$).

For the IIR DF shown in Fig. 1, the corresponding relations are as follows:

$$\begin{cases} b_0 = -t_{54}, \\ b_1 = t_{21}t_{54} + t_{32}t_{54} - t_{53}, \\ b_2 = t_{21}t_{53} + t_{31}t_{54} - t_{52}, \\ b_3 = -t_{51}, \\ a_1 = -t_{21} - t_{32} - t_{43}, \\ a_2 = t_{21}t_{43} - t_{31} - t_{42}, \\ a_3 = -t_{41}. \end{cases} \quad (5)$$

Let us turn attention to the fact that the coefficients of the transfer function polynomials are the sum of some products of some DF coefficients.

III. Structural Precision of IIR DF

By the structural precision of the IIR DF we mean some characteristic that depends on the structural scheme, which describes the accuracy of calculations invariantly to the specific filter parameters (numerical values of the filter coefficients). In [8] it was shown that structural accuracy is determined by two factors - the degree of algebraic numbers, which are zeros and poles of the DF, and a factor, which we will call the structure of the transfer function coefficients.

A. Algebraic Number Nature of Zeros and Poles of Practicable IIR DF is the First Constituent Unit of Structural Precision

The coefficients of any practically implementable digital filter have a finite bit depth, so they are elements of the set of rational numbers. From this it follows that the coefficients b_i and a_i of the numerator and denominator polynomials of the transfer function (4) for performable IIR DF do not belong to the continuum of real numbers, as it would be without taking into consideration the FWL of the filter coefficients. As the sum of products of rational numbers, they are also rational numbers.

At the same time, it is known [15] that the roots of polynomials with rational coefficients are elements of a countable set of algebraic numbers. Thus, the zeros and poles of practically realizable IIR DF are algebraic numbers. Algebraic numbers are characterized by a degree (below we will talk about the degree of zeros and poles). The degree of an algebraic number is the degree of the minimal (canonical) polynomial of this number - the only polynomial of the smallest degree with the leading coefficient equal to one [15].

The maximum degree of an algebraic number that is a root of a polynomial of the n^{th} degree with rational coefficients is equal to the degree of this polynomial. But in the general case, the degree of the root of the polynomial can be less than n [16]. In [14], [18] the analysis of the degrees of zeros and poles of some classical structures of IIR DF was carried out. The degree of zeros and poles of direct forms (I, II, transposed forms) is equal to the order of the DF. The degree of the zeros and poles of the cascade form of even order is two. The degree of the poles of the parallel form of an even order is two, and the degree of zeros equal the order of the filter.

Studies of the properties of a topological matrix [16] show that when generating topological matrices, the degree of poles can be controlled. The degree of the poles is related to the "fine" structure of the topological matrix in the following way.

In the matrix $\mathbf{T}(z)$ it is possible to distinguish square submatrices $\mathbf{T}_i(z) = \left[T_{i,k,l}\right]$, with dimension N_i, such that $T_{i,1,N_i} = z^{-1}$ ($\mathbf{T}_1$, $\mathbf{T}_2$, $\mathbf{T}_3$ and $\mathbf{T}_4$ in Fig. 2).

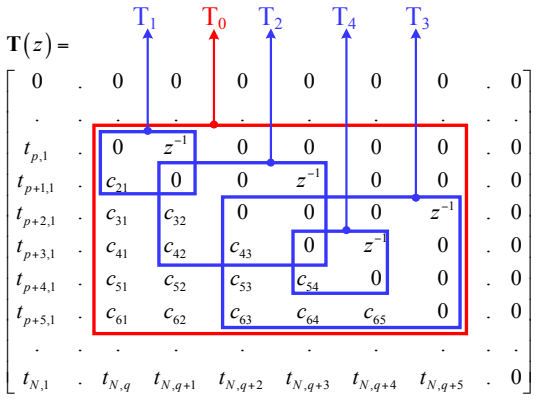


Fig. 2. An example of a "fine" structure of a topological matrix

Next, we single out the submatrix $\mathbf{T}_0$, which includes such submatrices $\mathbf{T}_i$ so that each submatrix $\mathbf{T}_i$ has at least one common element with another submatrix of the matrix $\mathbf{T}_0$.

The described operation of forming the submatrix $\mathbf{T}_0$ is denoted as

$$\mathbf{T}_0 = \prod_{i=1}^{n} \mathbf{T}_i . \tag{6}$$

The submatrix $\mathbf{T}_0$ described by (6) defines n poles of the n^{th} degree. These poles are independent of elements that are not elements of $\mathbf{T}_0$. Thus, Fig. 2 describes the IIR DF, which has four poles, which are algebraic numbers of the fourth degree. Unfortunately, for zeros, such a simple approach to the formation of the degree cannot be obtained. So, when generating a topological matrix, you can provide the required degree of poles and zeros. Let us now show that the degree of the poles is one of the factors of structural precision.

With the same digit capacity of the fractional part of the polynomial coefficients, the power of the countable set of admissible roots increases rapidly with an increase in the degree of algebraic numbers (Fig. 3 and Fig. 4), which means that the accuracy of the implementation of the poles increases.

So, the vector

$$\mathbf{algpw} = \left[n_z \quad n_p\right], \tag{7}$$

where n_z and n_p are degrees of zeros and poles, respectively, is the first constituent unit of IIR digital filter structural precision.

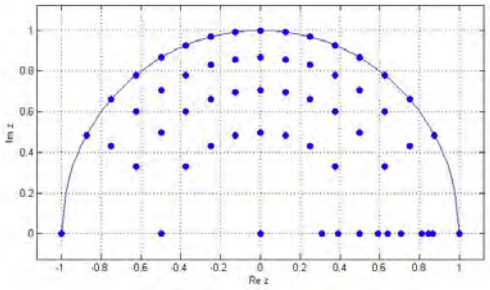


Fig. 3. A subset of algebraic numbers z of the second degree (n=2, $|z| \le 1$, $\mathrm{Im}\, z \ge 0$, digit capacity of the fractional part of the minimum polynomial coefficients m = 2)

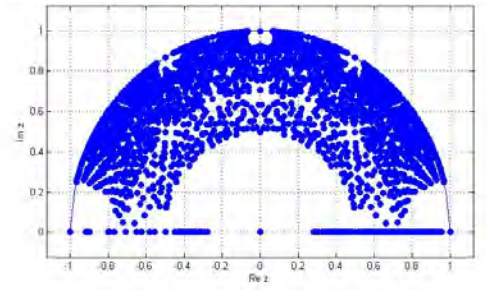


Fig. 4. A subset of algebraic numbers z of the fourth degree (n=4, $|z| \le 1$, $\mathrm{Im}\, z \ge 0$, digit capacity of the fractional part of the minimum polynomial coefficients m = 2)

B. Relationship Between the Fractional Part Length of Factors and Products

Obviously, the operation of multiplication leads to an increase in the length of the fractional part of the product compared to the digit capacity of the fractional part of the factors. Indeed, if s_i is the digit capacity of the fractional part of f_i, then the length of the fractional part of the product

$$P = \prod_{i=1}^{p} f_i \tag{8}$$

is equal to

$$m_P = \sum_{i=1}^{p} m_i . \tag{9}$$

For the above example of a digital filter, the coefficients of which are determined by the system of equations (5), assuming that the width of the fractional parts of all coefficients are the same and equal to m, we can write expressions for the bit depth of the fractional parts of the filter coefficients:

$$\begin{cases} m_{b_0} = m, \\ m_{b_1} = 2m, \\ m_{b_2} = 2m, \\ m_{b_3} = m, \\ m_{a_1} = m, \\ m_{a_2} = 2m, \\ m_{a_3} = m. \end{cases} \tag{10}$$

C. *Structure of Coefficients of Transfer Function Is the Second Constituent Unit of Structural Precision*

The second unit of structural precision is a parameter that we call "coefficient structure". It consists of two components. First component

$$\mathbf{scn} = \left[s_{b_0} \quad \dots \quad s_{b_n} \right] = \frac{1}{m} \left[m_{b_0} \quad \dots \quad m_{b_n} \right] \qquad (11)$$

is a vector whose elements are the structures of the coefficients of the transfer function numerator.

The second component is defined as follows:

$$\mathbf{scd} = \left[s_{a_1} \quad \dots \quad s_{a_n} \right] = \frac{1}{m} \left[m_{a_1} \quad \dots \quad m_{a_n} \right]. \qquad (12)$$

The nature of these two parts is completely different. The second component was studied in [13]. In this work, we investigate **scn**.

Taking into account (10), we can write down the following relations

$$\mathbf{scn} = \frac{1}{m} \left[m \quad 2m \quad 2m \quad m \right] = \left[1 \quad 2 \quad 2 \quad 1 \right]. \qquad (13)$$

With the same length of the fractional parts of the transfer function coefficients, the structure of the coefficients is a vector, the elements of which are the maximum number of factors in the products that make up the expressions of the coefficients of the transfer function through the coefficients of the filter structural scheme.

IV. Generation Structural Diagrams of a Predetermined Structure of the Coefficients of the Transfer Function Numerator

In this section, for various topological matrices, expressions for the structures of the transfer function numerator coefficients were obtained when performing symbolic calculations in the Maple system.

Obviously, this way is very inconvenient for the practical use of the developed approach to the synthesis of IIR DF with FWL. It has established the following aim. If we do not completely abandon the use of symbolic calculations, then at least significantly reduce their volume. Attempts were made to identify patterns in the formation of the structure of coefficients.

When generating structures, the following sequence of actions is assumed.

First, a submatrix $\mathbf{T}_0$ is generated, which provides a given degree of poles, a maximum degree of zeros, and a given structure of the coefficients of the transfer function denominator [13]. Rows and columns are then added to the submatrix $\mathbf{T}_0$ to form a matrix $\mathbf{T}(z)$ that provides the desired structure of the transfer function numerator coefficients.

The process of analyzing the structures of the coefficients corresponding to IIR DF with a topological matrix $\mathbf{T}(z)$ and having different numbers of input and output nodes is accompanied by the construction of matrices $\mathbf{S}$, $\mathbf{S}_0$, and $\mathbf{S}_i$, isomorphic to the matrices $\mathbf{T}(z)$, $\mathbf{T}_0$ and $\mathbf{T}_i$, respectively.

The process of forming the matrix **T** is shown in Fig. 5. The following symbols are used in this and subsequent figures.

- The boundaries of the submatrix $\mathbf{S}_0$.
- The boundaries of the submatrix $\mathbf{S}_i$.
- The boundaries of the matrix $\mathbf{S}$.
- [ ] - S_{ij} = **scn** is element of the matrix $\mathbf{S} = \left[S_{ij} \right]$. i and j are the numbers of the output and input nodes of the IIR DF.
- [ ] - Diagonal element of matrix $\mathbf{S}$.

Above and to the right of the matrix $\mathbf{S}$, the numbers of rows and columns are shown taking into account the change in the dimension of the matrix $\mathbf{S}$.

In Fig. 6 - 8 examples of the formation of matrices **S** for different matrices $\mathbf{S}_0$ corresponding to the poles of the second, third and fourth degrees are shown

1	2	3	4	5	6	7	8	9	10				
	1	2	3	4	5	6	7	8	9				
		1	2	3	4	5	6	7	8				
			1	2	3	4	5	6	7				
[C 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]				
[1 0 0]	[C 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]				1
[2 0 0]	[1 0 0]	[C 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]	[0 0 0]			1	2
[3 5 5]	[2 4 4]	[1 3 3]	[C 1 0]	[0 2 2]	[0 1 1]	[0 C 0]	[0 0 C]	[0 0 0]	[0 0 0]	1	2	3	
[4 5 5]	[3 4 4]	[2 3 3]	[1 2 0]	[C 2 2]	[0 2 2]	[0 1 0]	[0 0 1]	[0 0 0]	[0 0 0]	2	3	4	
[5 6 5]	[4 5 4]	[3 4 3]	[2 3 0]	[1 2 2]	[C 2 2]	[0 2 0]	[0 0 2]	[0 0 0]	[0 0 0]	3	4	5	
[6 6 0]	[5 5 0]	[4 4 0]	[3 3 0]	[2 2 0]	[1 1 0]	[C 0 0]	[0 C 0]	[0 0 0]	[0 0 0]	4	5	6	
[7 6 0]	[6 5 0]	[5 4 0]	[4 0 0]	[3 3 0]	[2 3 0]	[1 3 0]	[C 3 0]	[0 0 0]	[0 0 0]	5	6	7	
[8 7 6]	[7 6 5]	[6 5 4]	[5 4 0]	[4 4 3]	[3 4 3]	[2 4 0]	[1 4 3]	[C 0 0]	[0 0 0]	6	7	8	
[9 8 7]	[8 7 6]	[7 6 5]	[6 5 0]	[5 5 4]	[4 5 4]	[3 5 0]	[2 5 4]	[1 0 0]	[C 0 0]	7	8	9	
[10 9 8]	[9 8 7]	[8 7 6]	[7 6 0]	[6 6 5]	[5 6 5]	[4 6 0]	[3 6 5]	[2 0 0]	[1 0 0]	8	9	10	

Fig. 5. The process of forming the matrix **S**

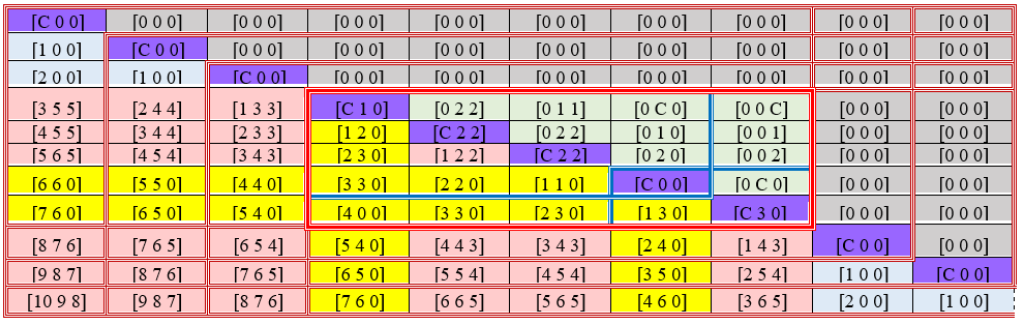


Fig. 6. Example of matrices **S** of a filter with two poles of the second degree

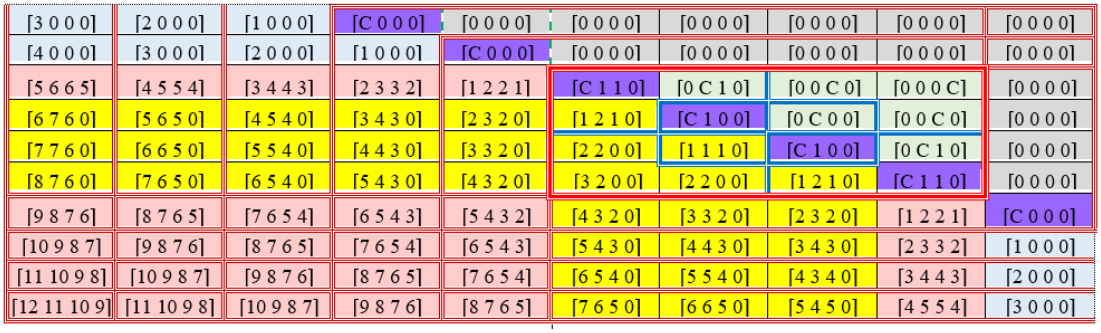


Fig. 7. Example of matrices **S** of a filter with three poles of the third degree

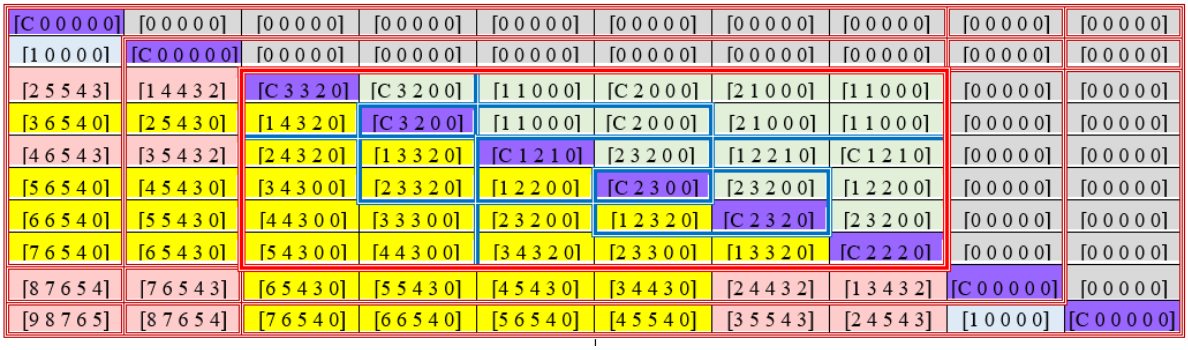


Fig. 8. Example of matrices **S** of a filter with four poles of the fourth degree

In this and subsequent figures the following conventions are used.

- Cell type A=Q (None of the vector **scn** components are zero).
- Cell type A=Y (At least one of the components of the vector **scn** is equal to zero)
- Elements of the submatrix $\mathbf{S}_0$ located above the main diagonal
- The numerator of the IIR CF transfer function is equal to a constant that is not equal to zero.
- The numerator, and hence the IIR DF transfer function, is zero.

It is obvious that such input-output pairs are of interest, which correspond to structures of coefficients, all of whose components are not equal to zero (Fig. 9). These pairs correspond to the numerators of the transfer function with n+1 degrees of freedom. In the figures, these options correspond to cells of type A=R.

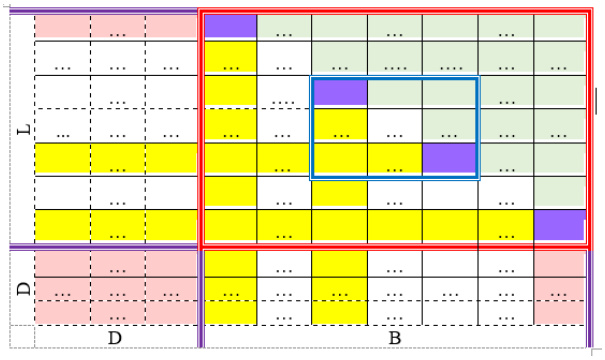


Fig. 9. Matrix **S** area of interest

Analysis of the structures shows that cells of the type A=Q are located in four regions of the matrix **S**.

1) *Region* **L**. The numbers of the first and last rows of this region are equal respectively to the numbers of the first and last rows of the submatrix $\mathbf{S_0}$. The region columns are located to the left of the columns of the submatrix $\mathbf{S_0}$.

2) *Region* **D**. The rows of this region are located below the rows of region **L**. The columns of this region are located to the left of the columns of the submatrix $\mathbf{S_0}$.

3) *Region* **B**. The numbers of the first and last columns of this region are equal respectively to the numbers of the first and last columns of the submatrix $\mathbf{S_0}$. The rows of this region are located below the lines of the submatrix $\mathbf{S_0}$.

4) *Region* **R**. The cells of this area are elements of the submatrix $\mathbf{S_0}$ and are located below its main diagonal.

The analysis of the structures revealed the following patterns.

- All cells of region **D** are of type A=Q.
- In the rows of region **L** with numbers equal to the numbers of the last rows of any submatrix $\mathbf{S}_i$, all cells are of type A=Y.
- The last row of the submatrix $\mathbf{S_0}$ is necessarily the last row of any submatrix $\mathbf{S}_i$. Therefore, any cell of the last row of region L has type A=Y.
- In the columns of region **B** with numbers equal to the numbers of the first columns of any submatrix $\mathbf{S}_i$, all cells are of type A=Y.
- All cells of regions **L**, **B** and **R**, which are not previously defined cells of type A=Y, are cells of type A=Q.
- The first row of the submatrix $\mathbf{S_0}$ cannot be the last row of any submatrix $\mathbf{S}_i$, therefore all cells of the first row of the area **L** are cells of the type A=Q.
- The last column of the submatrix $\mathbf{S}_0$ cannot be the first row of any submatrix $\mathbf{S}_i$, therefore all cells of the last column of the region **B** are of type A=Q.

It is very simple, regardless of n and N_0, to form the values of the components of the vectors **scn** in the region **D** (Fig. 10).

[N_0+3 N_0+2 N_0+1 … N_0-n+1]	[N_0+2 N_0+1 N_0 … N_0-n]	[N_0+1 N_0 N_0-1 … N_0-n-1]
[N_0+4 N_0+3 N_0+2 … N_0-n+2]	[N_0+3 N_0+2 N_0+1 … N_0-n+1]	[N_0+2 N_0+1 N_0 … N_0-n]
[N_0+5 N_0+4 N_0+3 … N_0-n+3]	[N_0+4 N_0+3 N_0+2 … N_0-n+2]	[N_0+3 N_0+2 N_0+1 … N_0-n+1]

Fig. 10. Values of **scn** in cells of region **D**

The values of the **scn** components in regions **L** and **B** are closely related to the filling of the cells of region **D**.

For any n, the values s_{b_0} are calculated as shown in Fig. 11.

j+1	j	…	2	1	C	0	…	0
j+2	j+1	…	3	2	1	C	…	0
j+3	j+2	…	4	3	2	1	…	0
…	…	…	…	…	…	…	…	…
j+i+1	j+i	…	i+2	i+1	i	i-1	…	0
…	…	…	…	…	…	…	…	…
N_0+j	N_0+j-1	…	N_0+1	N_0	N_0-1	N_0-2	…	C
j+N_0+1	j+N_0	…	N_0+2	N_0+1	N_0	N_0-1	…	1
j+N_0+2	j+N_0+1	…	N_0+3	N_0+2	N_0+1	N_0	…	2

Fig. 11. s_{b_0} values

Fig. 12 shows the process of calculating the s_{b_n} components

N_0-n+3	N_0-n+2	N_0-n+1	N_0-n	0					
N_0-n+3	N_0-n+2	N_0-n+1	N_0-n	0					
0	0	0	0	0					
0	0	0	0	0	0	0		0	
N_0-n+4	N_0-n+3	N_0-n+2	N_0-n+1	0	0	N_0-n		N_0-n	
N_0-n+5	N_0-n+4	N_0-n+3	N_0-n+2	0	0	N_0-n+1		N_0-n+1	

Fig. 12. s_{b_n} values

The remaining elements of the **scn** vector also depend on the structure of the submatrix $\mathbf{S}_0$.

To determine the elements $s_{b_{n-1}}$ of cells of type A=Q in the region **L**, we define the submatrix $\hat{\mathbf{S}}_0$. It is formed in the same way as the submatrix $\mathbf{S}_0$, only when it is formed, the submatrix $\mathbf{S}_1$ is not taken into account (Fig. 13):

$$\hat{\mathbf{S}}_1 = \prod_{i=2}^{n} \mathbf{S}_i . \qquad (14)$$

Let for the cell located in the first row and in the last column of the region **D**, the equality

$$s_{b_{n-1}} = N_0 - n + 2 = u \qquad (15)$$

is true.

Split region **L** into two subregions $\mathbf{L}_b$ and $\mathbf{L}_u$. The row numbers of the $\mathbf{L}_b$ subregion are equal to the row numbers of the $\hat{\mathbf{S}}_1$ submatrix. The rows of the subregion $\mathbf{L}_u$ are located above the rows of the submatrix $\hat{\mathbf{S}}_1$. All elements $s_{b_{n-1}}$ of cells of type A=Q in the last column of the $\mathbf{L}_b$ subregion are equal

$$s_{b_{n-1}} = u - 1 = N_0 - n + 1 . \qquad (16)$$

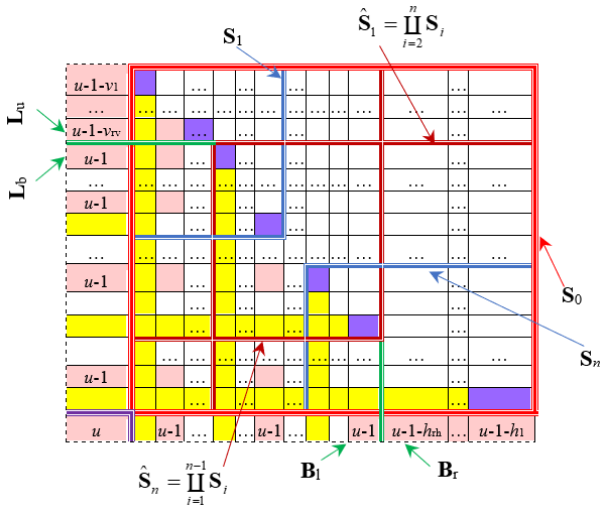


Fig. 13. $s_{b_{n-1}}$ values

All cells of the $\mathbf{L}_u$ subregion are of type A=Q. The subregion $\mathbf{L}_u$ consists of r_v rows. The value of r_v depends on the position of the submatrix $\mathbf{S}_2$ in the submatrix $\mathbf{S}_0$. For n=2 r_v=N_0-N_2. In the last column of the subregion $\mathbf{L}_u$, $s_{b_{n-1}}$ is described by the equality

$$s_{b_{n-1}} = u - 1 - v_i = N_0 - n + 1 - v_i, \tag{17}$$

where i=1, …, r_v, i=1 corresponds the upper row of subregion $\mathbf{L}_u$,

$$v_i = \begin{cases} 0, & \text{if} \quad i = rv, \\ 1, & \text{if} \quad i < rv. \end{cases} \tag{18}$$

The elements **scn** of cells of type A=Q in region **B** are calculated according to a similar principle. The submatrix $\hat{\mathbf{S}}_0$ is defined as follows

$$\hat{\mathbf{S}}_n = \prod_{i=1}^{n-1} \mathbf{S}_i . \tag{19}$$

Region **B** is divided into two subregions $\mathbf{B}_l$ and $\mathbf{B}_r$. The columns of the subregion $\mathbf{B}_r$ are located to the right of the columns of the submatrix $\hat{\mathbf{S}}_n$. All elements $s_{b_{n-1}}$ of cells of type A=Q in the first row of the $\mathbf{B}_l$ subregion are equal

$$s_{b_{n-1}} = u - 1 = N_0 - n + 1. \tag{20}$$

All cells of the $\mathbf{B}_r$ subregion are of type A=Q. The subregion $\mathbf{B}_r$ consists of r_h columns. The value of r_h depends on the position of the submatrix $\mathbf{S}_{n-1}$ in the submatrix $\mathbf{S}_0$. For n=2 r_h=N_0-N_1. In the first row of the subregion $\mathbf{B}_r$, $s_{b_{n-1}}$ is described by the equality

$$s_{b_{n-1}} = u - 1 - h_i = N_0 - n + 1 - h_i, \tag{21}$$

where i=1, …, r_h, i=1 corresponds the last column of subregion $\mathbf{B}_r$,

$$h_i = \begin{cases} 0, & \text{if} \quad i = rh, \\ 1, & \text{if} \quad i < rh. \end{cases} \tag{22}$$

Above, the calculation of the components of the **scn** vectors was considered only for the last column of region **A** and for the first row of region **B**. Fig. 14 and Fig. 15 explain the calculation of the remaining vectors of these regions.

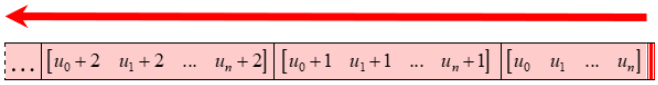


Fig. 14. The relationship between the cells of the rows of the region **L**.

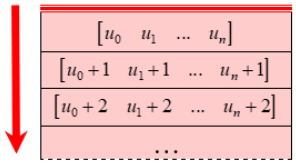


Fig. 15. The relationship between the cells of the columns of the region **B**.

Unfortunately, we were unable to obtain general analytical expressions for all components of the vectors of the last column of region **A** and the first row of region **B**. However, it seems that from the point of view of obtaining the maximum precision of the transfer function coefficients, the DF structure described by the topological submatrix of the type shown in Fig. 16 is of interest. To the previously calculated submatrix $\mathbf{T}_0$, a column (or columns) of region **A** and a row (or rows) of region **B** are added. We select the first and the last as the input and output nodes of the DF, respectively. The structure of the coefficients of the numerator of the transfer function in this case will be described by the cell of the region **D**, located in the first column and the last row of the submatrix **S**. However, other cells of type A=Q may be preferable from the point of view of the rounding noise level.

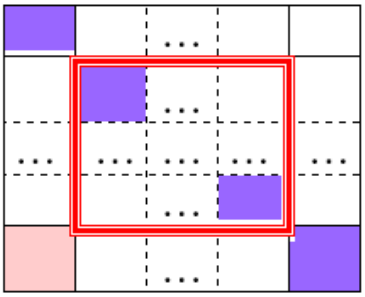


Fig. 16. Proposed topological submatrix structure

CONCLUSIONS

This work continues the series of publications in which a new paradigm for the synthesis of recursive FWL filters is proposed. Using the number-theoretic approach, already at the stage of functional synthesis, consideration is given to the finite bit depth of the transfer function coefficients. In the next step, a matrix description of the structure is generated, providing an accurate implementation of the previously calculated zeros and poles. To reduce the dimension of the problem, it is proposed to consider the so-called structural accuracy of filters, which does

not depend on its specific parameters. The structure of the transfer function coefficients introduced by us is one of the factors of structural accuracy. It turns out that for the numerator and denominator of the transfer function, the structure of the coefficients is determined in a completely different way. This factor was previously investigated for the denominator. In this work, heuristically obtained results for the numerator.

REFERENCES

[1] C. Rader, "dsp History –The rise and fall of recursive digital filters," *IEEE Signal Processing Magazine*, vol. 23, no. 6, pp. 46–49, 2006.

[2] "Does Altera support IIR compiler?," URL: https://www.intel.com/content/www/us/en/programmable/support/support-resources/knowledge-base/solutions/rd04072011_238.html

[3] W. Hess, *Digitale Filter: eine Einführung*, Springer Fachmedien Wiesbaden GmbH, 1993.

[4] A. Antoniou, *Digital Signal Processing: Signals, Systems, and Filters*, New York: McGraw-Hill, 2006.

[5] D. Schlichthärle, *Digital Filters: Basics and Design*, Berlin, Heidelberg: Springer-Verlag, 2011.

[6] V. Lesnikov, and T. Naumovich, "Number-theoretic and algebraic aspects of structural synthesis of digital filters," *GSPx-2004, The International Embedded Solutions Event (The Embedded Signal Processing Conference)*, Santa Clara, Ca, USA, September 27 - 30, 2004, paper number: 1374.

[7] 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.

[8] V. Lesnikov, T. Naumovich, and A. Chastikov, "Synthesis of recursive digital filters with finite word length: problems and their solutions," *Problems of Advanced Micro- and Nanoelectronic Systems Development*, 2019, Issue III, Moscow, IPPM RAS, pp. 46-53. Available: http://www.mes-conference.ru/books/proc-MES-2019-p3.pdf.

[9] 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.

[10] V. Lesnikov, T. Naumovich, A. Chastikov, and A. Metelyov, "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," *IEEE East-West Design and Test Symposium (EWDTS 2019)*, Batumi, Georgia, 13-16 Sept., 2019.

[11] V. Lesnikov, and T. Naumovich, "Generation and enumeration of structures of IIR digital filters," *GSPx-2005, Pervassive Signal Processing (The Embedded Signal Processing Conference)*, Santa Clara, Ca, USA, October 24 – 27, 2005, ISBN 0-9728718-2-9, paper number: 1837.

[12] V. Lesnikov, T. Naumovich, and A. Chastikov, "Generation and decomposition of digital filter decomposition," *IEEE East-West Design and Test Symposium (EWDTS 2017)*, Novi Sad, Serbia, 29 Sept. - 2 Oct., 2017.

[13] V. Lesnikov, T. Naumovich, and A. Chastikov, "Implementation of the poles of IIR digital filters with a declared structural precision," *30th International Conference Radioelektronika (RADIOELEKTRONIKA 2020)*, Bratislava, Slovakia, 15-16 April, 2020.

[14] R. E. Crochier, and A. V. Oppenheim, "Analysis of linear digital circuits," *Proceedings of IEEE*, vol. 63, no 4, 1975, pp. 581 – 595.

[15] D. Hilbert, *The theory of algebraic number fields*, Berlin – Heidelberg – New York: Springer – Verlag, 1998.

[16] 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.

[17] 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.

[18] V. Lesnikov, T. Naumovich, A. Chastikov, and A. Metelyov, "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," *IEEE East-West Design and Test Symposium (EWDTS 2019)*, Batumi, Georgia, 13-16 Sept. 2019, 6 p.

Set-membership Sparsity-Aware Proportionate Normalized Least Mean Square Algorithms for Active Noise Control

Felix Albu
Department of Electronics
Valahia University of Targoviste
Targoviste, Romania
felix.albu@valahia.ro

***Abstract*—Several set-membership (SM) based proportionate normalized least-mean-square (PNLMS) using the modified filtered-x (MFx) structures are proposed for active noise control. It is shown by simulations that the proposed algorithms that have a reduced numerical complexity can obtain better performance than MFx-PNLMS for various ANC paths.**

Keywords—active noise control, adaptive algorithms, set-membership

I. INTRODUCTION

Active noise control (ANC) is a technique for removing noise from a system by creating an anti-noise using an adaptive filter [1]-[3]. The modified filtered-x (MFx) approach [4] has a good convergence speed although it has a higher numerical complexity than the filtered-x approach. Many MFx-based schemes have been proposed (e.g. [5] - [12] and the references therein).

It is known that the system to be identified in ANC applications can have a certain degree of sparsity [1], [3]. The proportionate algorithms can achieve good performance on sparse system identification applications from various areas such as echo cancellation, acoustic feedback cancellation for hearing aids (e.g. [13]-[18] and the references therein). The proportionate algorithms proved better convergence for sparse primary or secondary ANC paths ([7]-[9], [11], and [19]). These alternatives involving the proportionate normalized least mean square algorithm (PNLMS) and the reweighted zero attracting (ZA) principle, called reweighted zero attracting PNLMS (RZA-PNLMS). The l_p-norm-constrained proportionate normalized least-mean-square (LP-PNLMS) algorithm and its version for ANC systems have been proposed too [11], but the choice of their parameter is complicate.

The set-membership (SM) filtering techniques have been proposed in order to reduce the numerical complexity and improve the estimation performance of the adaptive algorithms [19]. The SM filtering technique utilizes a bound on the magnitude of the estimation error and split the adaptive filtering algorithms into one step of information evaluation and another one of parameter update [19]. The reduction of complexity is obtained by updating less frequently the filter weights. By combining the set-membership and proportionate principles, the set-membership PNLMS was introduced [20].

In [21], the use of the correntropy induced metric (CIM) approach has led to numerous CIM-based adaptive algorithms. In [20] the correntropy induced metric (CIM) penalized SM-PNLMS algorithm was proposed and its use for acoustic channel estimation was investigated. The cost function of the set-membership PNLMS (SMPNLMS) included the CIM penalty [20].

In this paper a novel approach to incorporate the set membership principle to several PNLMS algorithms and integrate them into a MFx structure is presented by investigating the mean-square deviation (MSD) convergence curves for various sparseness primary and secondary paths. The application of the proposed set-membership based algorithms for ANC is novel and has not been investigated yet.

Section II presents the used ANC system and the update equations of three proposed algorithms: the MFx set-membership PNLMS (MFx-SMPNLMS); the MFx reweighted zero attracting set-membership PNLMS (MFx-RZASMPNLMS) and the MFx correntropy induced metric set-membership PNLMS (MFx-CIMSMPNLMS). In Section III, the performance comparison of the proposed algorithms is presented. Section IV concludes the paper and suggest directions for further study.

II. THE PROPOSED ALGORITHMS

In broadband feedforward ANC, the noise is reduced by subtracting from the acoustic signal a generated signal by using an error signal [1]. In the MFx structure shown in Fig. 1, $\mathbf{q}(k)$ is the primary plant, the instantaneous error signal $\hat{e}(k)$ is estimated [7]. The signal $x(k)$ is filtered using an estimation of the secondary path, $\hat{\mathbf{s}}(k)$ generating the filtered signal, $x_f(k)$. In this structure it is not needed for the signal $d(k)=\mathbf{x}(k)\mathbf{h}(k+1)$ to be available, where $\mathbf{x}(k)=[x(k)\ x(k-1)\ \ldots x(k-L)]$, and L is the filter length [11]. The condition $\hat{d}(k)=\mathbf{x}_f(k)\mathbf{h}(k+1)$ is imposed, where $\mathbf{x}_f(k)=[x_f(k)\ x_f(k-1)\ldots x_f(k-L)]$, $\hat{y}(k)=\mathbf{x}_f(k)\mathbf{h}(k)$ $\hat{e}(k)=\hat{d}(k)-\hat{y}(k)$ [7]. More details about the MFx structure can be found in [4] and [7].

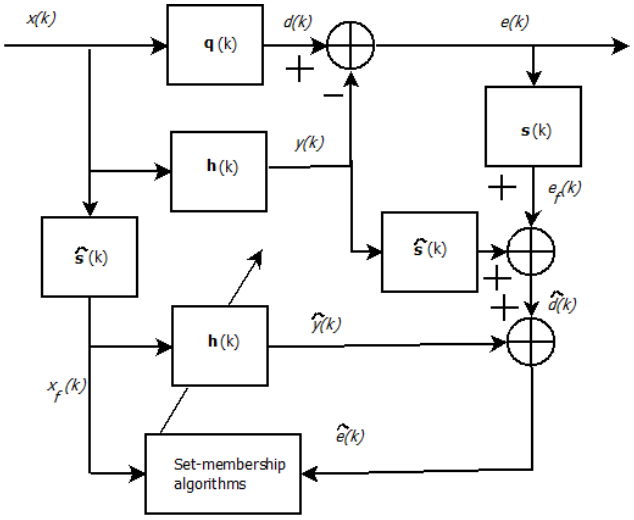


Fig. 1. The proposed set-membership based algorithms approach for ANC.

A. The MFx-SMPNLMS algorithm

In the PNLMS algorithm the larger taps converge faster than the smaller taps [13]. The MFx-SMPNLMS algorithm update equation is the following:

$$\mathbf{h}(k+1)=\mathbf{h}(k)+\mu_{SM}(k)\frac{\hat{e}(k)\mathbf{G}(k)\mathbf{x}_f(k)}{\mathbf{x}_f^T(k)\mathbf{G}(k)\mathbf{x}_f(k)+\delta_{PNLMS}}, \quad (1)$$

where $\delta_{PNLMS}=\sigma_{x_f}^2/L$ is a regularization factor, the gain matrix $\mathbf{G}(k)$ is [11]

$$G(k)=diag\left(g_0(k),g_1(k),...,g_{L-1}(k)\right), \quad (2)$$

with each $g_i(k)$ computed as follows

$$g_i(k)=\frac{\chi_i(k)}{\sum_{i=0}^{L-1}\chi_i(k)} \qquad 0\le i\le L-1, \quad (3)$$

with

$$\chi_i(k)=\max\left[\rho_g \ \max\left[\delta_p,|h_0(k)|,...,|h_{L-1}(k)|,|h_i(k)|\right]\right], \quad (4)$$

where ρ_g and δ_p are small positive constants [11]. The step size is $\mu_{SM}(k)$

$$\mu_{SM}(k)=\begin{cases}1-\gamma/|e(k)|, & \text{if } |e(k)|>\gamma\\ 0, & \text{otherwise.}\end{cases} \quad (5)$$

When $\mu_{SM}(k)=\mu$ has a fixed value, the MFx-PNLMS algorithm is obtained.

B. The MFx-RZASMPNLMS algorithm

The MFx reweighted zero attraction Set-Membership PNLMS algorithm (MFx-RZASMPNLMS) can be easily obtained by using the same gain matrix $\mathbf{G}(k)$, regularization factor and step size as above. The updating function of MFx-RZASMPNLMS algorithm that incorporate the RZA and proportionate principles [7] is:

$$\mathbf{h}(k+1)=\mathbf{h}(k)+\mu_{SM}(k)\frac{\hat{e}(k)\mathbf{G}(k)\mathbf{x}_f(k)}{\mathbf{x}_f^T(k)\mathbf{G}(k)\mathbf{x}_f(k)+\delta_{PNLMS}} - \rho_{RZA}\frac{\text{sgn}(\mathbf{h}(k))}{1+e_p|\mathbf{h}(k)|}, \quad (6)$$

where ρ_{RZA} is the zero attracting strength parameter, e_p is the shrinkage magnitude parameter [7] and $\text{sgn}(\bullet)$ is the signum function.

By comparing (Eq. 1) with (Eq. 6) it can be easily observed that the MFx-ZASMPNLMS has $3L$ more multiplications than the MFx-SMPNLMS algorithm (it includes multiplications with the inverse of the denominator values and other element-wise operations from the third term of (Eq. 6)).

C. The MFx-CIMSMPNLMS algorithm

If the CIM penalty is integrated with the SMPNLMS algorithm the CIMSMPNLMS algorithm is obtained [20]. Following the same approach from [20] and combining with integrating sparsity algorithms from [7], [8] the update equation from (Eq. 7) is obtained for the MFx implementation, called Modified Filtered-X Correntropy Induced Metric Set-Membership PNLMS (MFx-CIMSMPNLMS) algorithm. The third term additional term from (Eq. 7) is the CIM zero attractor that is influenced by the ρ_{CIM} and the kernel width, σ.

By comparing (Eq. 1) with (Eq. 7) it can be easily observed that the MFx-CIMSMPNLMS has $3L$ more multiplications than the MFx-SMPNLMS algorithm (it includes the element-wise operations from the third term of (Eq. 7)) and L exponential function computation. Due to this fact, the MFx-CIMSMPNLMS algorithm is also the most complex among the proposed set-membership based algorithms. As will be shown in the next section the percentage of weight updates can vary between 2% and 53% depending on the sparseness of the primary plant. Therefore, for these percentage values, the proposed algorithms can have a much lower numerical complexity than the MFx-PNLMS algorithm (that has 100% updates of the second term in (Eq. 1), (Eq. 6) and Eq. (7). Depending on the system sparseness, these computational savings on the third term can be higher than the computation of the third term of the same corresponding update equations.

$$\mathbf{h}(k+1)=\mathbf{h}(k)+\mu_{SM}(k)\frac{\hat{e}(k)\mathbf{G}(k)\mathbf{x}_f(k)}{\mathbf{x}_f^T(k)\mathbf{G}(k)\mathbf{x}_f(k)+\delta_{PNLMS}} - \rho_{CIM}\frac{1}{L\sigma^3\sqrt{2\pi}}\mathbf{h}(k)\exp\left(-\frac{(\mathbf{h}(k))^2}{2\sigma^2}\right). \quad (7)$$

III. Simulation Results

This section compares the results of simulations of the MFx-PNLMS, MFx-SMPNLMS, MFx-RZASMPNLMS, and MFx-

CIMSMPNLMS algorithms for an ANC application using the same primary plants and secondary paths as in [7]. For each primary path (sparse (density 1/800), partially sparse (density 73/800) and non-sparse (density 785/800)), the algorithms were run for 45,000 iterations with the secondary path set as sparse (density 1/800) at the start of the experiment, changed to partially sparse plant (density 73/800) after 15,000 iterations and to a non-sparse path (density 785/800) after additional 15,000 iterations. For the simulations the following parameters were used: $e_p = 0.05$, $\delta_p = 0.0000625$, $\rho_g = 0.01$, $\sigma = 0.01$, $\gamma = 0.0447$ [19], $\mu = 0.1$ and the length of all the paths and filters is L=800. The performance of the algorithms has been figured by the MSD convergence curves.

It is obvious that the parameter ρ_{RZA} of the MFx-RZASMPNLMS algorithm and the parameter ρ_{CIM} of the MFx-CIMSMPNLMS algorithm have to be carefully chosen for the best performance and balance the convergence speed and steady-state performance for plants with various sparseness values. The ρ_{RZA} and ρ_{CIM} values were selected from five equally spaced values between 10^{-7} and 10^{-5} and the average MSD value for the last 2500 iterations for each secondary sparse level was computed. The results for ρ_{RZA} and ρ_{CIM} are shown in Table I and Table II, respectively.

The best ρ_{RZA} and ρ_{CIM} values from Table I and Table II were used for the MFx-RZASMPNLMS and MFx-CIMSMPNLMS algorithms, respectively. Figures 2-4 shows a comparison of the performances of the MFx-PNLMS, MFx-SMPNLMS, MFx-RZASMPNLMS and MFx-CIMSMPNLMS algorithms for a sparse, partially sparse and non-sparse plants, respectively.

It can be noticed from Fig. 2 that the proposed set-membership versions do not obtain good performance in the case of sparse plant and sparse and partially sparse secondary paths, the best algorithm being the MFx-PNLMS. Only for non-sparse secondary path, the MFx-CIMSMPNLMS algorithm performs better than the competing algorithms. In these cases, the number of updates is less than 4% if compared with the MFx-PNLMS algorithm (see Table III). The number of updates performed by the set-membership based algorithms in cases depicted in Figs. 2-4 is shown in Table III.

From Fig. 3 it can be noticed that the MFx-SMPNLMS obtains the best performance in the case of partially sparse plant and non-sparse and partially sparse secondary paths.

TABLE I. AVERAGE MSD VALUE FOR THE MFX-RZASMPNLMS ALGORITHM FOR VARIOUS ρ_{RZA} VALUES

Primary plant	**Average MSD values for the last 2500 iterations (best ρ_{RZA} *value*)**		
Sparse	-49.68 (10^{-5})	-30.38 (10^{-5})	-37.97 (10^{-5})
Partially sparse	-31.47 (10^{-7})	-15.51 (10^{-6})	-16.93 ($5\cdot10^{-7}$)
Non sparse	-28.78 (10^{-7})	-10.86 (10^{-7})	-10.87 (10^{-7})

TABLE II. AVERAGE MSD VALUE FOR THE MFX-CIMSMPNLMS ALGORITHM FOR VARIOUS ρ_{CIM} VALUES

Primary plant	**Average MSD values for the last 2500 iterations (best ρ_{CIM} *value*)**		
Sparse	-43.38 (10^{-5})	-24.16 (10^{-5})	-27.94 ($5\cdot10^{-6}$)
Partially sparse	-30.84 ($5\cdot10^{-7}$)	-15.00 (10^{-6})	-15.58 (10^{-7})
Non sparse	-28.46 (10^{-7})	-11.06 (10^{-7})	-9.30 (10^{-7})

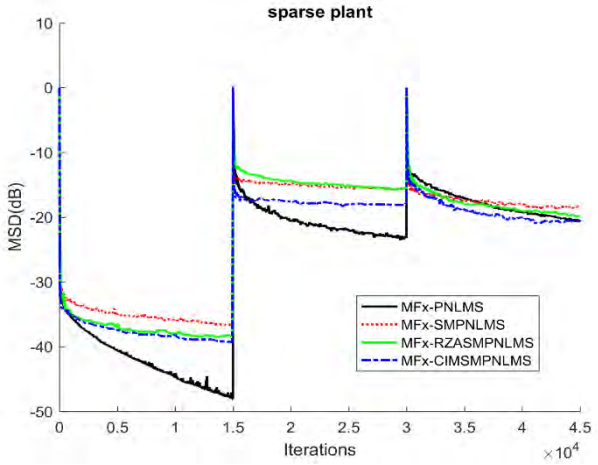


Fig. 2. The MSD performance of the investigated algorithms for a sparse plant.

In these cases, the number of updates is less than 12% if compared with the MFx-PNLMS algorithm (see Table III). For the sparse secondary path and partially sparse plant, the MFx-PNLMS algorithm has a slower initial convergence speed but achieves a lower MSD value than competing set-membership based algorithms.

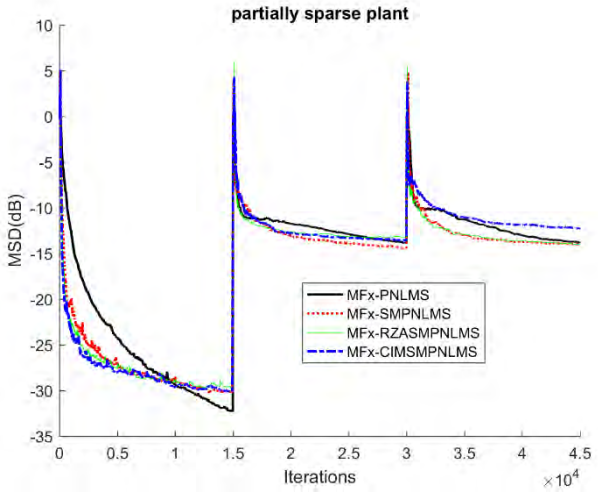


Fig. 3. The MSD performance of the investigated algorithms for a partially sparse plant.

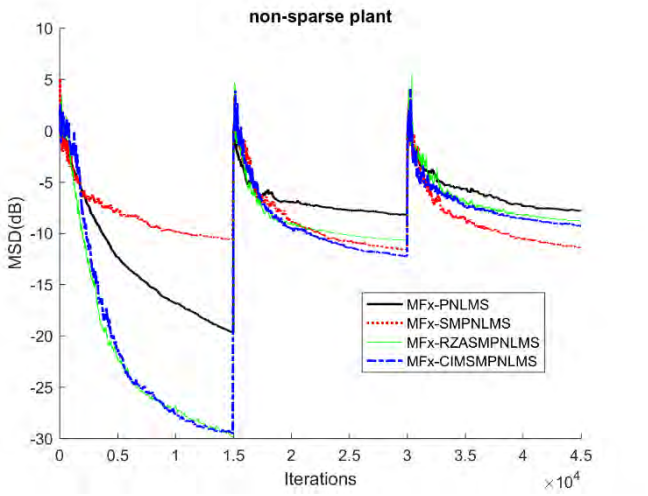


Fig. 4. The MSD performance of the investigated algorithms for a non sparse plant.

TABLE III. PERCENTAGE OF THE UPDATES FOR THE CONSIDERED SET-MEMBERSHIP ALGORITHMS

Primary plant	Percentage of updates		
	MFx-SMPNLMS	***MFx-RZASMPNLMS***	***MFx-CIMSMPNLMS***
Sparse	3.75	3.82	2.97
Partially sparse	11.36	9.59	10.31
Non sparse	51.52	50.27	51.93

In the case of using a non-sparse plant (Fig. 4), the MFx-RZASMPNLMS and MFx-CIMSMPNLMS have the best MSD performance for the sparse secondary path, the MFx-CIMSMPNLMS has the best performance for the partially sparse secondary path and MFx-SMPNLMS is the best for non-sparse secondary path. However, for this case, the number of updates ranges from 50 to 52% (see Table III). Therefore, it can be concluded that there is a connection between the sparseness of the primary path and the number of updates performed by the proposed set-membership based algorithms.The complexity reduction is higher for sparse primary plants than for non-sparse primary plants. From practical point of view, the partially sparse and non-sparse cases are relevant, the sparse cases are rather extreme cases for ANC systems.

IV. CONCLUSIONS

This paper has proposed three new set-membership based algorithms using the PNLMS algorithm for ANC. The simulation results demonstrate that the proposed algorithms can provide MSD improvements over the MFx-PNLMS algorithm for semi-sparse and non-sparse plants and secondary paths. The main advantage of the reduced average numerical complexity is emphasized and therefore, the considered set-membership based algorithms can represent a good choice for practical ANC systems. In the future work the corresponding variable step-size versions MFx or filtered-x structures will be investigated.

REFERENCES

[1] S. M. Kuo and D. R. Morgan, Active Noise Control Systems: Algorithms and DSP Implementations, John Wiley and Sons Inc., New York, NY, 1996.

[2] V. I. Djigan, A. A. Petrovsky, J. Qin, Y. Song, "Modified hybrid active noise control system," in Proc. of 2015 IEEE East-West Design & Test Symposium (EWDTS).

[3] Y. Kajikawa, W. S. Gan and S. M. Kuo, "Recent advances on active noise control: open issues and innovative applications," APSIPA Trans. Sig. Inf. Process., vol. 1, pp. 1-21, Aug. 2012.

[4] E. Bjarnason, "Active noise cancellation using a modified form of the filtered-x LMS algorithm," in Proc. 6th Eur. Signal Process. Conf., Brussels, Belgium, 1992, pp. 1053–1056.

[5] A. Gonzalez, F. Albu, M. Ferrer, and M. de Diego, "Evolutionary and variable step size strategies for multichannel filtered-x affine projection algorithms," IET Signal Process., vol. 7, no. 6, pp. 471–476, Aug. 2013.

[6] M. T. Akhtar, M. Abe, and M. Kawamata, "Modified-filtered-x LMS algorithm based active noise control system with improved online secondary-path modeling," in Proc. of MWSCAS2004, Hiroshima, Japan, Jul. 25–28, pp. I-13–I-16, 2004.

[7] A. Gully, R. C. de Lamare, "Sparsity aware filtered-x affine projection algorithms for active noise control", in Proc. of ICASSP 2014, pp. 6707-6711, 2014.

[8] F. Albu, A. Gully and Rodrigo C. de Lamare, "Sparsity-aware pseudo affine projection algorithm for active noise control," Asia-Pacific Signal and Information Processing Association, pp.1-5, 2014.

[9] F. Albu, Y. Li, Y. Wang, "Low-complexity non-uniform penalized affine projection algorithms for active noise control ", in Proc. of Eusipco 2017, Kos, Greece, pp. 1315-1319.

[10] J. G. A. Ochoa, G. S. Rivera, A. R. Silva, J. M. Guevara and G. A. Arzate, "Multichannel filtered-x set-membership affine projection-like algorithm," IEEE Latin America Transactions, vol. 16, no. 8, pp. 2131-2137, Aug. 2018.

[11] F. Albu, I. Caciula, Y. Li, Y. Wang, "The lp-norm proportionate normalized least mean square algorithm for active noise control", in Proc. of ICSTCC 2017, Sinaia, Romania, pp. 396-400.

[12] F. Albu, "The Constrained Stability Least Mean Square Algorithm for Active Noise Control", in Proc. of BLACKSEACOM 2018, Batumi, Georgia.

[13] D. L. Duttweiler, "Proportionate normalized least-mean squares adaptation in echo cancelers," IEEE Trans. Speech Audio Process., vol. 8, no. 5, pp. 508–518, Sep. 2000.

[14] Y. Li and M. Hamamura, "An improved proportionate normalized least-mean-square algorithm for broadband multipath channel estimation," The Scientific World Journal, vol. 2014, Article ID 572969, 9 pages, 2014.

[15] F. Albu, C. Paleologu, J. Benesty, and S. Ciochina, A low complexity proportionate affine projection algorithm for echo cancellation, in Proc. EUSIPCO, August 2010, pp. 6-10, 2010.

[16] Y. Li, Y. Wang, R. Yang, et al. "A soft parameter function penalized normalized maximum correntropy criterion algorithm for sparse system identification", Entropy 2017, 19(1), 45, 2017.

[17] F. Albu, C. R. C. Nakagawa, and S. Nordholm "Proportionate algorithms for two-microphone active feedback cancellation", in Proc. of EUSIPCO 2015, pp. 290-294.

[18] F. Albu, P. S. R. Diniz, "Improved set-membership partial-update pseudo affine projection algorithm", in Proc. of IEEE ICCACI 2016.

[19] S. Werner and P. S. R. Diniz, "Set-membership affine projection algorithm," IEEE Signal Processing Letters, vol. 8, no. 8, pp. 231–235, August 2001.

[20] Z. Jin, Y. Li, and Y. Wang, "An enhanced set-membership PNLMS algorithm with a correntropy induced metric constraint for acoustic channel estimation," Entropy, vol. 19, no. 6, Jun.2017.

[21] A. Singh and J.C. Principe, "Using correntropy as a cost function in linear adaptive filters," Proc. of 335 International Joint Conference on Neural Networks, pp.2950-2955, 2009.

Modelling Error Pulses in a CMOS Triple Majority Gate while Exposed to an Ionizing Particle

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*— Modelling results for noise pulses forming by logical elements are presented when collecting charge from the single particle tracks with wide range of the linear energy transfer of 10–90 MeV·cm^2/mg. This performs using 3D CAD physical models of CMOS transistors designed on 65 nm bulk technology with shallow trench isolation of transistor groups. The main positive thing that happens when collecting the charge at the linear energy transfer more 40 MeV·cm^2/mg from the track in the group of NMOS transistors belonging to element OR and in the group of PMOS transistors belonging to element AND is holding these transistors in the mode when all transistors are collecting charges. This forms a delay of noise pulses on the outputs of elements and decreases its duration.**

Keywords—charge collection, logical element, modelling, noise pulse, single particle, track

I. Introduction

CMOS combinational logic elements are devoted to simulation the impacts of single ionizing particles using physics-based device models, both two- and three-dimensional. It was noted [1] that the noise immunity of CMOS logic designed using bulk technology would decrease to the values of linear energy transfer by a particle on a track equal to 2 MeV·cm^2/mg when technology node will shrink to 100 nm or less. The increasing the duration of noise (error) pulses to 300–500 ps at the linear energy transfer (LET) value of 30 MeV·cm^2/mg also predicted [2].

At technology nodes below 100 nm the CMOS logic shows the influence of diffusion transfer of charge carriers induced on the same track on adjacent circuit nodes. This joint charge collection can lead to a reduction in the duration of noise pulse [3], known in the literature as "pulse quenching". In [4], we present the results of the study of CMOS transistors groups in AND and OR elements that are most sensitive to impacts of single ionizing particles in a triple majority gate.

The purpose of this work is to study effects of minimizing the noise (error) pulses while a charge collection from a track of a single nuclear particle. It may be in cases a charge collection by CMOS transistors of NAND and inverter elements, as well as NOR and inverter, made in a limited volume of silicon, surrounded by shallow trench isolation, for a wide range of the linear energy transfer by a particle to the track.

II. Triple Majority Gate on AND and OR Elements

Fig. 1 presents a diagram of a triple majority gate on three logical elements AND (D1-D3) and one element OR (D4). The circuit of the element AND on Fig. 2a consists of a NAND element and an inverter. The circuit of an element OR on Fig. 2b includes of an element NOR and an inverter.

The simulation of impacts of single nuclear particles on CMOS elements (designed on the bulk 65-nm CMOS technology) carried out using 3-D TCAD transistors models of the work [5]. Layouts of the topology of elements AND and OR are given in Fig. 3. The elements AND and OR consist each of two groups. Groups of NMOS transistors marked as Gr1N or Gr4N, groups of PMOS transistors marked as Gr1P or Gr4P

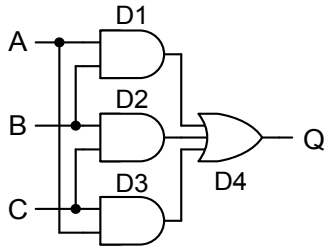
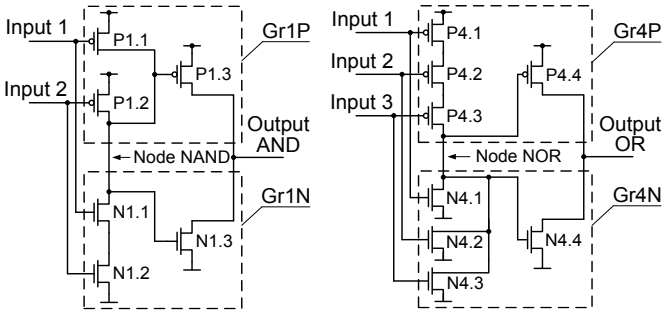


Fig. 1 Diagram of the triple majority gate based on AND (D1-D3) and OR elements (D4).



(a) (b)

Fig. 2. Circuits of the logical elements 2AND (a) and 3OR (b).

Funding: The reported study was funded by Russian Foundation for Basic Research, project number 19-07-00651

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

(Fig. 2, Fig. 3). The transistor widths are 400 nm in the AND element and 800 nm are in the OR element.

Groups of transistors in AND and OR elements surrounded by a shallow trench isolation of the 400 nm depth. The silicon groups Gr1N and Gr1P have dimensions of 885×400×400 nm^3, and the Gr4n and Gr4P groups – 1180×800×400 nm^3. Asterisk markers in Fig. 3 correspond to the input track points of single particles with the track direction normal to the crystal surface. Asterisks marked in red correspond to tracks that modeled with the range of linear energy transfer by particles to a track of 10–90 MeV·cm^2/mg. The results obtained using the hybrid TCAD-SPICE modeling, in which gates D1 (AND) and D4 (OR) are simulating by the 3D TCAD physical models, and two logic gates D2 (AND) are represented by the SPICE models.

As the result of the study [4], it was found that for tracks with LET = 60 MeV·cm^2/mg, noise pulses with the highest durations and amplitudes are formed in the group of NMOS transistors Gr4N of the OR element and the group of PMOS transistors Gr1P of the AND element. The biggest part of a charge from the track in these groups collected by transistors with a common drain area. In this paper, we study the parameters of noise pulses formed in the same groups, but at tracks with LET in the range of 10–90 MeV·cm^2/mg. Results obtained by Sentaurus Device at the temperature 25°C and the supply voltage of 1.0 V for particle tracks with linear transfer energy 10–90 MeV·cm^2/mg.

III. Features of a Charge Collection in an Element OR

Fig. 4 presents the forming of noise pulses in time at nodes of the element OR with input signals of the triple majority gate A = B = C = 0. Input track points are in the Gr4N group of NMOS transistors. In the case of an OR element with the input track point of 4n (Fig. 4a) at LET = 60 MeV·cm^2/mg, after switching the inverter during the formation of the track, the voltage on the drain of the NMOS transistor of the inverter decreases to a minimum $V_{MIN.OR.4n}$ when collecting the charge (electrons) from the particle track. This voltage level is fixed when it falls within the $V_{MIN.OR.4n}$ = +(0.02–0.7) V at LET ≥ 40 MeV·cm^2/mg (Fig. 4b). In this case, a "plateau" formed at the output OR with a slightly increasing voltage, which creates a delay in the formation of a positive polarity noise pulse and reduces its duration.

For the OR element with the input track point of 5n, the minimum voltage on its output while collecting electrons does not fall below $V_{MIN.OR.5n}$ = +(0.086–0.1) V even if LET ≥ 80 MeV·cm^2/mg (curves for 5n on Fig. 4b). In this case for the input track point of 5n, a "plateau" of a long-term non-increasing voltage does not form on the output OR, but immediately begins to form a positive polarity noise pulse.

The front of the positive polarity pulse at the output OR for the input track points of 4n and 5n is forming by the current of the open PMOS transistor of the inverter, charging the capacity of the output node OR when the NMOS transistor of the inverter is closed. In this case, the maximums of the noise pulses are reached at the output OR for the 4n and 5n input track points at the voltages at the NOR node in the range (0–0.1) V. The example on Fig. 4a demonstrates this fact for the input track points of 4n and 5n for LET = 60 MeV·cm^2/mg.

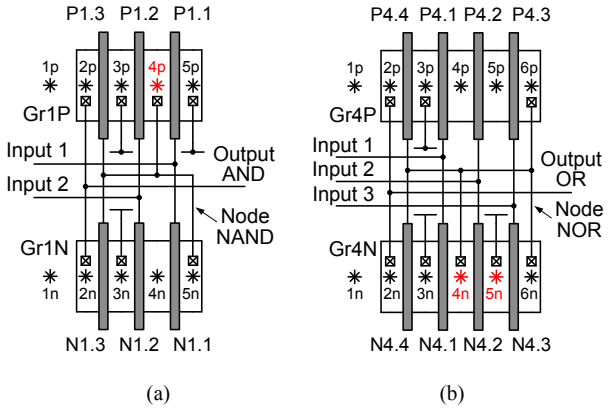
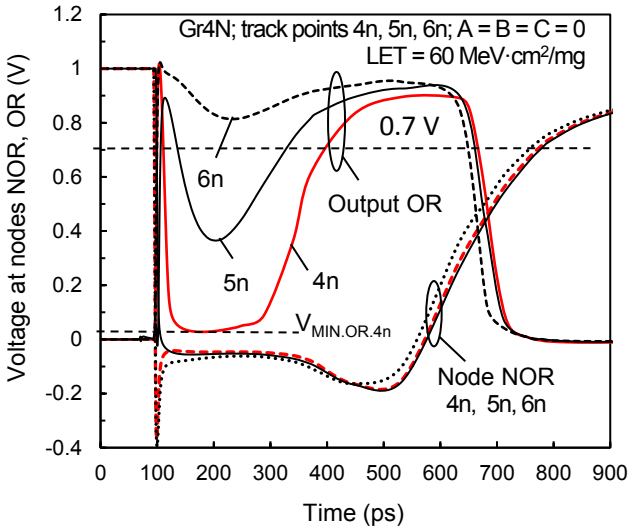


Fig. 3. Layouts of the logical elements 2AND (a) and 3OR (b).



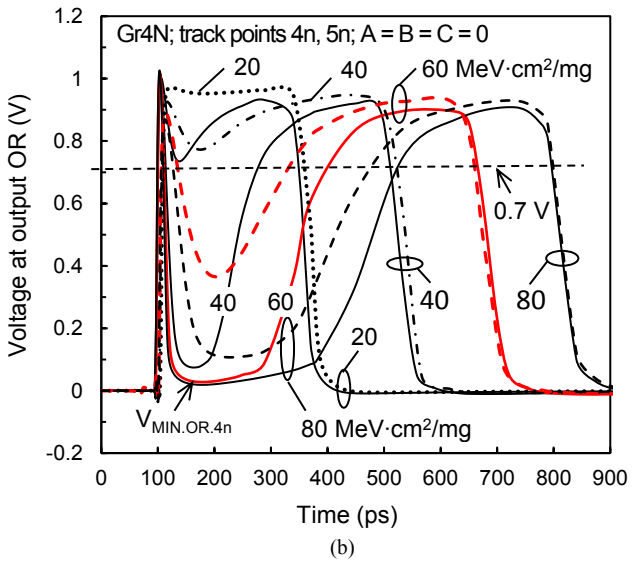


Fig. 4. Noise pulses at element nodes NOR at signals on the inputs of the majority element A = B = C = 0, the input track points in the Gr4N group: (a) track with LET = 60 MeV·cm^2/mg, the input points 4n, 5n, 6n; (b) tracks with LET = 20–80 MeV·cm^2/mg, the input points 4n, 5n.

The PMOS transistor of the inverter continues to charge the capacity of the output node OR after the noise pulse at the output OR exceeded the level of 0.7 V and the formation of the peak of the noise pulse begins. At this time, the voltage between its drain and source becomes less than 0.3 V, which translates this transistor from a flat to a steep area of the voltage characteristic, where the drain current decreases in proportion to the decrease in the voltage at the drain. This causes the charging capacity of the output node OR to slow down and the formation of the peak of the noise pulse to an amplitude value, which increases the duration of the interference pulse.

When the voltage at the NOR node and the input of the inverter exceeded 0.3 V the noise pulse reaches 0.7 V after the amplitude value, the inverter switched and the drop in the noise pulse increases sharply (Fig. 4a). Then the initial stationary state is restored at the output OR.

IV. Features of a Charge Collection in an Element AND

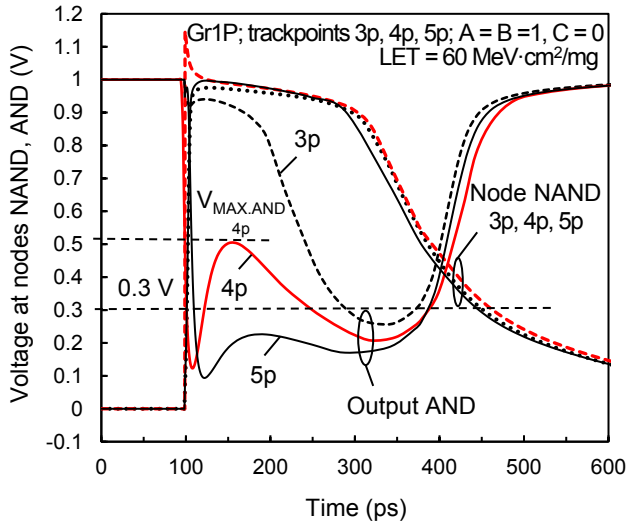
Fig. 5 presents the forming of noise pulses in time at nodes of the element AND with input signals of the triple majority gate A = B = 1, C = 0. Input track points are in the Gr1P group of PMOS transistors. When the inverter switches after forming a track, the closed PMOS transistor of the inverter begins to collect the charge (holes). It increases the voltage on the output AND forming a positive polarity pulse. The curves for the case of the track 4p shows on Fig. 5a. This voltage does not rise above $V_{MAX.AND}$ = + 0.74 V at the output AND even at LET = 90 MeV·cm^2/mg (Fig. 5b). This is not create a “plateau” of constant voltage. In this case immediately begins to form a noise pulse of the negative polarity. However, there is forming a delay of noise pulses at the level of 0.3 V decreasing of their duration at this level.

In the range LET = 10–90 MeV·cm^2/mg the amplitude values of the positive polarity pulses (Fig. b) at the output AND are reached when the node NAND voltages are in the range (0.75–0.85) V (Fig. 5a). In this mode, the PMOS transistor of this node is closed. After the noise pulse reaches the amplitude value, the inverter switching begins (Fig. 5a, 5b) and the noise pulse finishes sharply when the voltage at the input voltage of the inverter drops less the level 0.7 V. Then the initial stationary state is restored at the output AND.

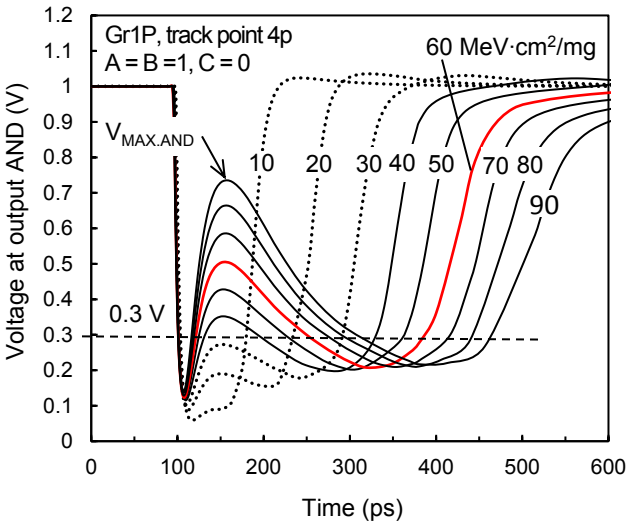
V. Analisis of Simulation Results

Fig. 6a shows the results of modeling of the pulse parameters of the element OR for the input track points 4n and 5n of the Gr4N group with inputs of the majority gate A = B = C = 0. Fig. 6a shows graphs illustrating the time parameters of noise pulses of element OR with input track points 4n and 5n depending on the linear energy transfer to the track in the range LET = 10–90 MeV·cm^2/mg. These curves include the duration of pulses at the node NOR $t_{PULSE.NOR}$ at the level of 0.3 V, the duration of noise pulses at the output OR $t_{PULSE.OR}$ at the level of 0.7 V and as well as the duration of delays in the rise of noise pulses to the level of 0.7 V at the output OR $t_{DEL.PULSE.OR}$.

The durations of the pulses on the node NOR at the input track points 4n and 5n are linear functions of the linear energy transfer. These dependences on Fig. 6a practically do not depend on the input points in the group Gr4N of NMOS transistors in the entire range LET = 10–90 MeV·cm^2/mg. They



(a)



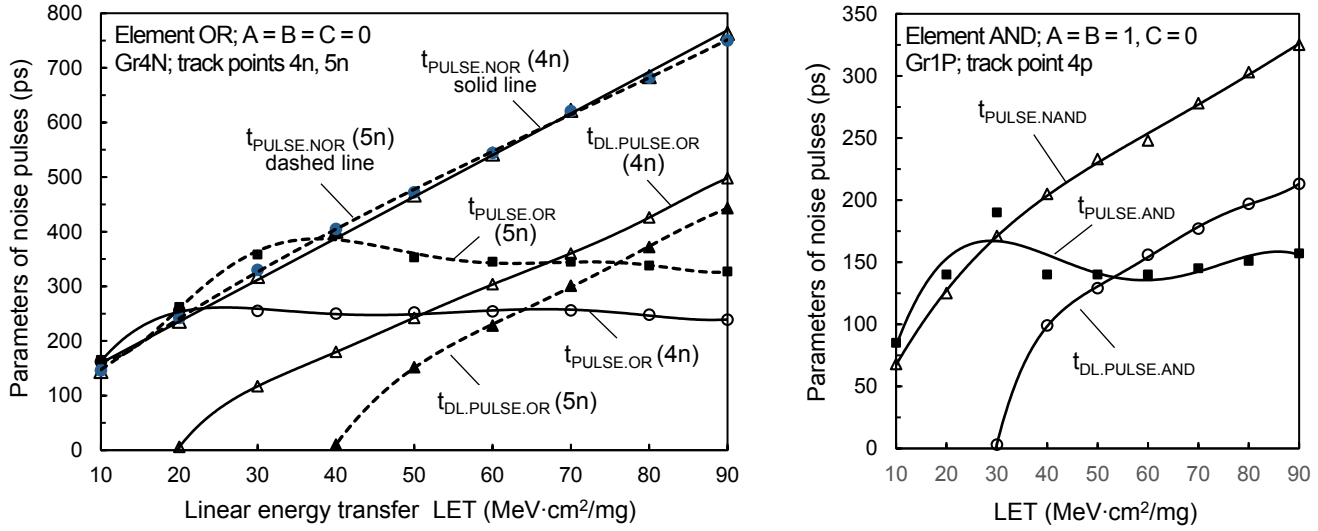
(b)

Fig. 5. Noise pulses at element nodes AND at signals on the inputs of the majority element A = B = 1, C = 0; the input track points in the Gr1P group: (a) track with LET = 60 MeV·cm^2/mg, the input points 3p, 4p, 5p; (b) tracks with LET = 20–90 MeV·cm^2/mg, the input point 4p.

characterize the total charge collection in the group Gr4N of NMOS transistors.

The time delays of noise pulses $t_{DELPULSE.OR}$ from a track impact to the noise pulse level of 0.7 V on the output OR are the linear functions of the values LET for the input track point 4n from 30 MeV·cm^2/mg and for 5n from 50 MeV·cm^2/mg (Fig. 6a). Time delays are more above 50–80 ps for the track point 4n due to the ”plateau" of the low level voltage in front of a noise pulse.

The duration of the noise pulses at the output OR (TMG output) $t_{PULSE.OR}$ virtually repeats the values at the node NOR $t_{PULSE.NOR}$ (Fig. 6a) with the input track point 4n from LET = 10 up to 20 MeV·cm^2/mg. For the input track point 5n this repeating is from LET = 10 up to 40 MeV·cm^2/mg. Up to LET = 90 MeV·cm^2/mg the duration of the noise pulses at the output OR repeated on the level 250 ps for the input track point 4n and



(a) (b)

Fig. 6. The time parameters of noise pulses depending on the linear energy transfer particle to the track: (a) the input points of 4n and 5n in the group of transistors Gr4N of the element OR; (b) the input point 4p of the group of transistors Gr1P of the element AND.

on the level 350 ps for the input track point 5n. The duration of the noise pulses at the output OR $t_{PULSE.OR}$ at the level of 0.7 V for the input track point 5n is 1.4–1.6 times more (Fig. 6a) in the range LET = 30–90 MeV·cm^2/mg of the duration of the noise pulses $t_{PULSE.OR}$ for the track point 4n.

Fig. 6b shows the modeling results of pulse parameters in the element AND for the input track point 4p of the Gr1P group, when majority gate inputs are A = B = 1; C = 0. These curves defined as functions of the linear energy transfer in the range from 10 to 90 MeV·cm^2/mg. In the case of the characteristics for the element AND, there is an almost linear dependence of noise pulse duration on the node NAND $t_{PULSE.NAND}$ at the level of 0.3 V as the LET function.

The duration of the noise pulses at the output AND virtually repeats the values at the node NAND (Fig. 6b) for LET = 10–30 MeV·cm^2/mg to the value of $t_{PULSE.AND}$ = 190 ps at LET = 30 MeV·cm^2/mg. For LET = 40–90 MeV·cm^2/mg the noise pulse durations fixed on the level $t_{PULSE.AND}$ = 140–150 ps due to common collecting the charge by all NMOS transistors of the Gr1P group. For the element AND, there is an almost linear dependence of the delay of the noise pulse rise $t_{DELPULSE.AND}$ to the level of 0.3 V as the function of LET in the interval 40–90 MeV·cm^2/mg. Two curves are virtually symmetrical: this are durations of the pulses at the node NAND $t_{PULSE.NAND}$ and delay durations of noise pulses on the output AND $t_{DL.PULSE.AND}$ with the time distance between them 100–110 ps.

The main thing that happens when collecting the charge from the track in the element OR with the track input points of 4n and 5n is holding the node NOR with a negative polarity voltage on the node, when the NMOS transistors of the node NOR are in the inverse mode offset. In this mode, the NMOS transistor of the inverter is closed and all NMOS transistors of the Gr4N group collect the charge (electrons) and transition it to the common bus of the element. This time interval is slightly less than the duration of the noise pulse detected at the node NOR at the 0.3 V level. At the final part of finding the NMOS transistors of the node NOR in the inverse mode offset, the noise pulse is formed, but not by increasing the charge collection, but by weakening its collection.

In case of the element AND, the output AND changes as the output of the element OR when a charge collects from the track with the input track point of 5n without the "plateau" of the low level voltage in front of a noise pulse. In both cases of OR and AND elements, the noise pulse forms in fact by an inverter, which returns the logical elements from a non-stationary state to the original stationary state without collecting a charge.

VI. Conclusion

The simulation showed the existence of reducing the noise pulse duration in logic elements made using CMOS technology with shallow trench isolation of transistor groups. The joint charge collection by transistors in different logical states reduces the noise pulse duration by two to five times due to increasing in the delay of its formation. The presented features of the elements is useful when designing CMOS microprocessor systems for space applications.

References

[1] P.E. Dodd, M.R. Shaneyfelt, J.A. Felix, and J.R. Shwank, "Production and propagation of single-event transients in high-speed digital logic ICs", *IEEE Transactions on Nuclear Science*, 2004, vol. 51, no. 6, pp. 3278–3284.

[2] P.E. Dodd, and L.W. Messengill, "Basic mechanisms and modeling of single-event upset in digital microelectronics", *IEEE Transactions on Nuclear Science*, 2003, vol. 50, no. 3, pp. 583–602.

[3] N.M. Atkinson, A.F. Witulski, W.T. Holman, J.R. Ahlbin, B.L. Bhuva, and L.W. Massengill, "Layout technique for single-event transient mitigation via pulse quenching", *IEEE Transactions on Nuclear Science*, 2011, vol. 58, no. 3. pp. 885–890.

[4] Yu.V. Katunin, and V.Ya. Stenin, "Modeling of single ionizing particles impact on logic elements of a CMOS triple majority gate", *Russian Microelectronics*, 2020, vol. 49, no. 3, pp. 214–223.

[5] 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.

Filtration of Diagnostic Data for Retrospective Analysis in Health Monitoring Systems of Engineering Structures

Dmitry V. Efanov,
D. Sc., Associate Professor,
First Deputy General Director –
Chief Engineer of Vega LLC,
Professor at Peter the Great St. Petersburg Polytechnic University
St. Petersburg, Russia
TrES-4b@yandex.ru

Valeriy Myachin,
D. Sc., Professor,
General Director of Scientific and Technical Center
"Integrated Monitoring Systems" LLC,
St. Petersburg, Russia
vmyachin@ipr.ru

German Osadchy,
Technical Director of Scientific and Technical Center
"Integrated Monitoring Systems" LLC,
St. Petersburg, Russia
osgerman@mail.ru

Marina Zueva,
Design engineer of Scientific and Technical Center
"Integrated Monitoring Systems" LLC,
St. Petersburg, Russia
marina-seo-media@yandex.ru

Abstract—Authors of this publication being discussed the task of diagnostic data filtration received from health monitoring systems of engineering structures. During operational period of conventional monitoring systems, initial diagnostic matrix contains as usual some of foreseen errors. Existence of those errors has various reasons, such as malfunction of data transmission form measuring devices, extraordinary data spikes, equipment failure, faults accumulated from extraneous signals, loss of data etc. Persistence of such kind of errors not just interdicts further expert analysis, but it blocks further solution regarding diagnostic with afterward forecast. Aforesaid circumstances drastically diminish the effectiveness of monitoring technology as well as spoils certainty result. In connection therewith, authors set a challenge concerning algorithms search (either combination of algorithms) to actualize initial data filtration operation. Russky Island Bridge in Vladivostok, Russia was designated as a sample of monitoring site, where started from 2012 permanent monitoring system was in-service. The entire data from the abovementioned system should be accumulated within temporary storage with afterward retrospective analysis. We applied such ways of static data filtration like time-series data, machine learning method plus implementation of embedded functions of information filtration. Be means of static methods completion we achieved the result of clear up to 50% of being diagnosed files. As for machine learning method it helped us to improve filtration function up to 60%. Passing through time-series method we made it about 70% of filtration quality. The best ever filtration outcome was shown by in Forecast Library via Artificial Programming Language (APL R) up to 98% of effectiveness, consequently that was the cause, why we apply this particular function for filtration performance of the initial diagnostic data nowadays. Filtered information further may be processed by machine analysis Artificial Intelligence (AI), which should ensure more precise estimation regarding service capability of unique structures.

Keywords—*health monitoring systems of engineering structures; filtration of diagnostic data; retrospective information analysis; control of pre-failure condition; service life prediction.*

I. INTRODUCTION

One of our challenging task concerning service and maintenance of complex technical structure, being consisted of hundreds and thousands interacted components with colossal dimensions, various forms and structures, plus number of elements, which have different types of loads to ensure it working capacity reliability, safety as well as fault tolerance on the highest level. Those types of structures are existing all over the world within various spheres, such as unique historical architecture, stadiums, skyscrapers, bridges, flyovers, motorways, railroads, etc. [1 – 8].

For construction sites preservation purposes technical maintenance, diagnostics and monitoring of technical status quo are being wide spread. [9]. Availability of up-to-date sensors of physical magnitudes, precise measuring gadgets with computer technologies for data accumulation, storage and processing of diagnostic information allows us to arrange systems of structured monitoring for sophisticated technical sites [10]. Implementation of the above systems not only simplifies the service and maintenance of technical objects, but it helps us to identify critical conditions of elements prior to failure status event, consequently, we may forecast possibility of trouble, estimate its service resource plus create those needed recommendations of proper maintenance measures.

The entire monitoring sets, regardless of relevant spheres, have similar structures as well as the same advantages and disadvantages. Experience of monitoring system performance in the sphere of transportation infrastructure showed us that data processing of received diagnostic information via correct interpretation is one of the keystone problems.

It has to be mentioned that modern monitoring systems can obtain huge volumes of diagnostic records even per single sensor only, let alone a whole kit of sensors. For example, if diagnostic features analysis of monitoring site is required (for instance, oscillation impact), 50 Hz polling rate of sensors is essential, which gives 180 thousands measurements per hour (equals 4.32 million measurements a day!).

978-1-7281-9899-6/20/$31.00 ©2020

Fig. 1. Monitoring site – Bridge to Russki Island, Russia.

Total volumes of outcome data for processing may achieve tens of gigabytes. Most monitoring systems are building up there information in temporary storages, from where they extract needed files to data base and the result of it is, as usual, trouble with diagnostic data synchronization. Even for this puzzle solution, mission of quality analysis of diagnostic info is the mater. Concerning diagnostic data status, we have frequently the following troubles: stop of data transmission from measuring device, unusual data spikes, equipment fault, errors with hindrances amassing, loss of data etc. Summarizing the above, for diagnostic operation and further mission fulfillment reckoning genesis plus forecasting, diagnostic information is considered as 'dirty data'. Subsequently the effectiveness of monitoring result is reckoned as low one.

One more of important monitoring task is that of inevitability of data processing compares to threshold limit values being designated by legitimate norms, rules and regulations, which excludes it quality analysis of the aforesaid information. For most assignments solutions concerning proper diagnostics of long term data files for tens of years are being claimed, which improves radical the results of monitoring conclusion. Information storages for long time periods must be provided with high reliability measures for data safety.

Consequently, initial data synchronization plus filtration of diagnostic info as well as valuable info regeneration in the event of lockdowns for afterward quality analysis via AI is the essential matter [11 – 17].

At the moment our designated mission is the search for universal option of raw diagnostic data filtration for structured systems of sophisticated engineering structures monitoring. We shall emphasize that in the sphere of monitoring diagnostics of technical structures the above mission is in fact actual one and 'ready-to-go solutions', in spite of monitoring challengers, are not exist yet.

II. Monitoring Site

Cable Stayed Bridge on Russky Island in Vladivostok, Russia was designated as a monitoring site by us. Above-mentioned structure (see Fig. 1) is reckoned as the longest bridge crossing in the whole world [18]. It was completed by the end of 2012. The site is located over the Bosporus Vostochny and links Nazimova Peninsula with cape Novosilskogo on Russki Island with world longest span of cable stayed structures of 1,104 meters long. Bridge height is 324 meters, which is considered as second one in the world. This unique structure is being presented on Russian bank paper of 2 000 rubles.

For bridge crossing technical status quo supervision during service period, structural monitoring system was arranged including 50 strain gauges with 88 sensors of strain with temperature control, 16 thermo sensors, 21 accelerometers, 12 two-coordinates inclinometers, two displacement sensors and six pressure probes. For the above data accumulation, 295 measuring channel were completed. On Fig. 2 you may see the screen shot out of monitoring performance with technological window of recorded info related to features per single pylon.

Data regarding measuring sensors are being stored on the Server of Database Management Interbase XE. Diagnostic information is being presented via tables, which were registered with frequency of one measurement per five seconds and single measurement per ten minutes. Accumulation of diagnostic features is being conducted permanent way via block by block sequences of data files, composed of sequences of designated period of group of channels readings. Permanent flow of data division per segments (either unification of instant readings into blocks) allows us to create working cycle of data processing, being activated parallel data buildup with little pause for algorithm processing. At the same time enduring block-by block sequences is the matter of total measured matrix ensurance without losses.

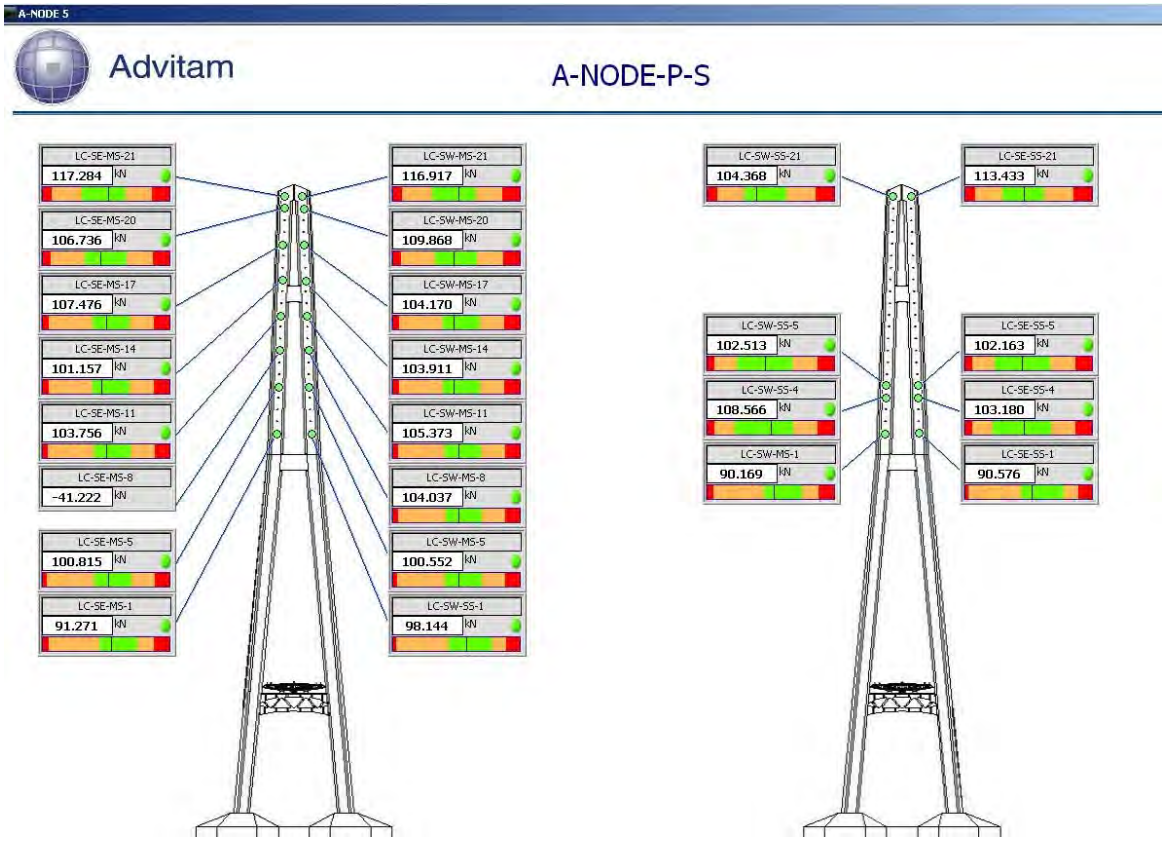


Fig. 2. Screen shot of technological window.

That dilemma of being received mass of data was in that spikes of faults being included into accumulated information, which cause can not be identified. Existance of those spikes was considered as a hindrance for analysis fulfillment. Other disadvantage is that within acquired mass of data there we almoust always have some spots of information absence for various reasons of monitoring such as electrical power blockout (routine outage, power loss etc.). Based on the above the task of filtration of diagnostic information from fault spikes with wild values is very important mission. Consequently, acquisition of 'clean data' (filtered) allows us to fulfull proper analysis of diagnostic info.

III. Experiments of Diagnostic Data Filtration

For the task completion regarding diagnostic data acquisition, which being useful for onward contemplation and processing together with AI application, algorithm otherwise combination of algorithms concerning array of raw data filtration from anomaly spikes ought to be discovered. For this purpose, authors of this article are being researched various methods of data filtration linked to real-time structured monitoring, being received form Russky Island, Vladivostok (RF).

During contemplation reckoning fair reasonable method of filtration being targeted diagnostic 'raw data', there were more than 500 matrixes of diagnostic info. Let us present you our experiments description relating data filtration from two sensors of diagnostic information, such as M6L2DP-strain gauge and RBG4DP – temperature sensor. Total accumulated by us data are being covered the period of single months specifically (March 2019). On Fig. 3 one may see 'raw diagnostic data', where you may notice some anomalies which looks like sharp shift of measured results. Experimental study of those types of diagrams considered impossible as well as AI method implementation.

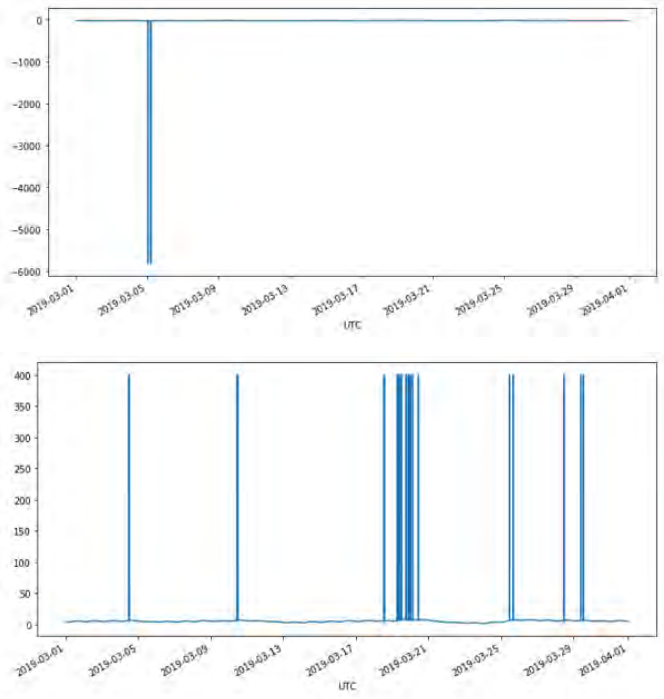


Fig. 3. 'Raw' data with anomalies (Python): upper diagram – data from M6L2DP-strain gauge, lower diagram – data from RBG4DP-temperature sensor.

Static methods of data filtration such as 'three sigma' rule and '99 Percentile' were initially tested by us [19].

Precise analysis of ‘three sigma’ rule demonstrated us some filtration quality improvement (Fig. 4). Hence, it is not working so immaculate way and we can observe some leftover ‘tails’, consequently we can not apply the above method to any other mass of diagnostic information files. As for ’99 Percentile’ approach it worked much better way for presented data, meanwhile it results annihilation of some info per extremum data (Fig. 5). For other mass of diagnostic data we watched absolutely opposite effects, thus somewhere the first method was better, somewhere the second one. But commonly we noted that straight implementation of static methods may not bring us to proper solution of diagnostic data filtration.

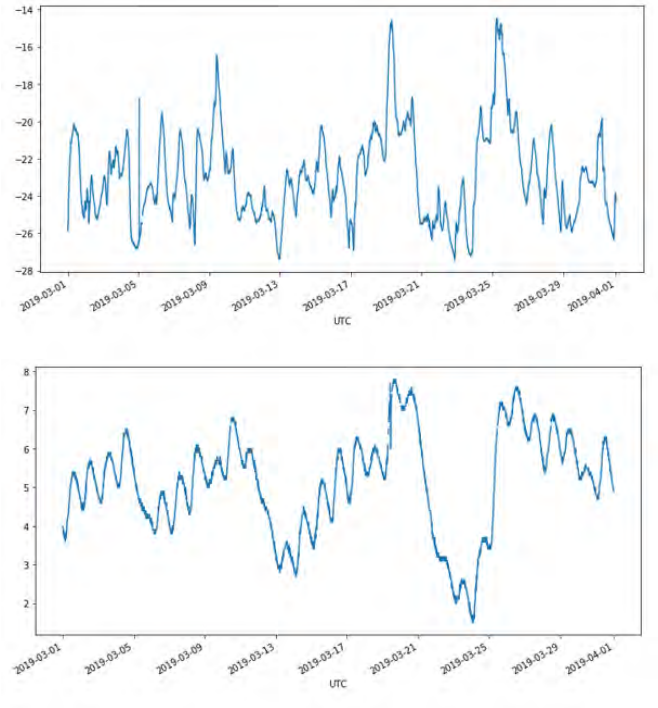


Fig 4. Filtered data from ‘three sigma method’ (Python): upper diagram – data from M6L2DP-strain gauge; lower diagram – data from RBG4DP-temperature sensor.

Functioning of static methods may be suitable only for data with few anomalies and certain allocation (close to normal distribution). For the entire arrays of diagnostic data static methods must not present quality end result, accordingly implementation of static methods considered as not universal, for the cause there is not a single one, which may be effectively applied to the entire matrix of diagnostic info.

Next, let us take a glance at machine learning methods, which quality looks much better than some arrays of static methods, but several ones appeared to be worse. Let us review, for instance, realization of one-class vehicle of Support Vector Machine (SVM) [20] for data from the same sensors.

During filtration of diagnostic data via One-class SVM standard characteristics of Python-Scikit-learn (sklearn.svm.OneClassSVM) [21] were being used. On (Fig. 6) filtration results of diagnostic info per RBG4DP–temperature sensor are being shown.

- Important steps for realization of sklearn.svm.OneClassSVM, considered the following features:
- kernel – (linear; poly; radial basic function (rbf); sigmoid; self designated);
- nu – upper bounder per errors percentage and lower bounder per percentage of reference vectors (0.5 natively);
- degree – grade of poly kernel;
- gamma –factor for kernel function (1/n_features natively);
- coef0 – feature per function of poly either sigmoid kernel.

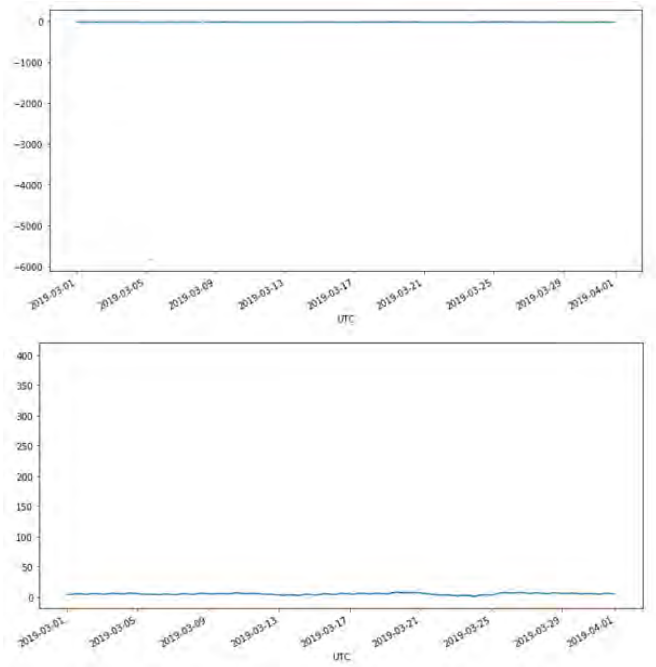


Fig. 5. Filtered data via ‘99Percentile’ (Python): upper diagram – data from M6L2DP – strain gauge; lower diagram – data from RBG4DP temperature sensor.

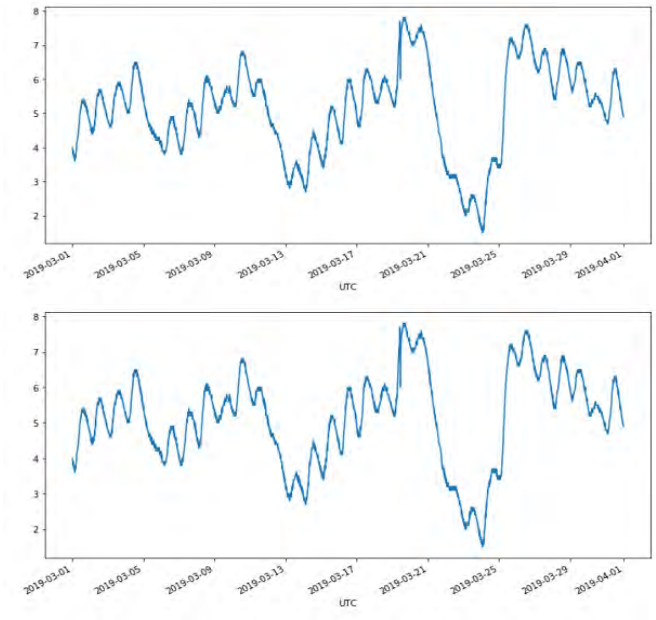


Fig. 6. Filtered data from RBG4DP temperature sensor via One-class SVM method (kernel – «rbf», methods from library Scikit-learn, Python): upper diagram – during normal features; lower diagram – if gamma=0.1.

Implementation of sklearn.svm.OneClassSVM, with designated characteristics for diagnostics data array resulted the fact that some of single-shot anomalies rested unnoticed. Such as, similar result was obtained out of M6L2DP-strain gauge. Based on the abovementioned some endeavors were tested for proper parameters of filter choice to improve the procedure of diagnostic 'raw data' filtration. For instance, we exchanged 'gamma' for designated model. Most suitable result was achieved by us via parameter gamma = 0.1. Augmentation of the above parameter effected vanishing of useful diagnostic data. We shall make mention, that OneClassSVM approach produced even worse outcome compares to 'three sigma rule'.

For experimental purposes other methods of machine learning was assigned. For instance, IsolationForest [22] (one of options per 'random forest' [23]). This approach considered trees arrangement up to the moment of data extraction. During tree branching arrangement random criteria should be chosen as well as random splitting (threshold of split). For each measured value designated criterion should be defined – arithmetical average of tree leaves depth, where it was isolated.

Important realization features of sklearn.ensemble.IsolationForest:

- n_estimatorsint – number of basic estimation per ensemble;
- max_samplesint – groups quantity, on which the split should be conducted;
- contamination – share of splits per max_samplesint – number of random certain data file;
- max_featuresint – criteria quantity, per which the split to be identified (in our case it is the single criterion).

Method 'IsolationForest' was implemented with the following features: contamination=0.1 and max.fearures=1.0. Filtration outcome with present option compares to 'One-class SVM' method is being shown on Fig. 7. As for 'IsolationForest' course, it is being annihilating a lot of useful diagnostic information. Amendments to software model characteristics did not bring positive results.

Later we tested 'time-series data' methods and the best of all were 'ETS' and 'ARIMA' [24].

'Model ETS' presented rather good results concerning data filtration from RBG4DP sensor Fig. 8 (upper diagrams). The idea of the above replica is exponential smoothing option (predictive method), (where value variable includes the entire previous periods into forecast and via exponential curve is being loosing self weight as time goes on). This model is completed by means of 'ETS. () function' from library fpp2 in terms of R-language.

'ARIMA' models – considered integrated models of autoregression (moving-average model). As for updated 'ARIMA', application of those models should not be conducted straight to designated 'time series', but initial differentiation is essential via acquisition of single 'time series' as a differences result of subsequent values of initial 'time series'. Filtration products by means of 'ARIMA' options are being shown on Fig. 8 (see lower diagrams).

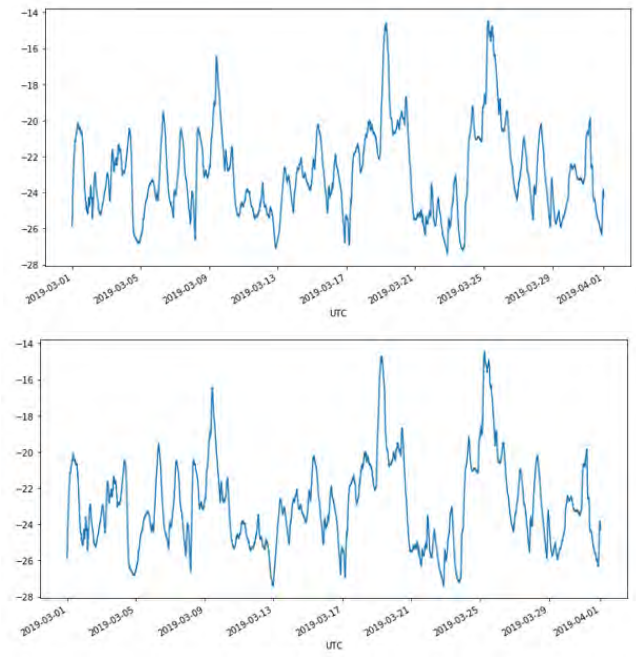


Fig. 7. Filtered data from M6L2DP sensor via 'machine method' (Python, methods from library Scikit-learn): upper diagram – by means of 'One-class SVM' (kernel – 'rbf') if standard features; lower diagram via 'IsolationForest' method.

Let us clear up that filtration performance via models 'ETS' & 'ARIMA'. On Fig. 8 you may observe the search sequence and anomalies vanishing with present models execution (upper diagrams match the purge of single data matrix from RBG4DP sensor through the 'ETS'-model, lower – data filtration from M6L2DP-sensor with an aid of 'ARIMA' method. On the left side you may notice initial diagrams with anomalies plus in the middle part are those illustrations of highlighted anomalies and on the right side 'cleared out' line charts. It should be pointed out, that in the event of 'ETS model' application to data of second sensor one of spikes remained the same.

Models for 'time-series' were enough successful for the task of diagnostic 'raw data' filtration. Hence, for the realization of aforesaid models, automated inspection procedure for which one to be suitable per this or that file must be inserted. Additionally one more checking operation should be added, like the one to analyze the necessity of any spike quest per certain file to avoid extra functions engagement, which is crucial for the reason that always amasses of valuable information extinction from database. For resolution of the before mentioned dilemma, we believe, may be tested hypothesis reckoning anomaly observation existence, for example by means of 'Irving method' [25]. Meanwhile, 'Irving method' has several limits, which not allows us to cleanse diagnostic data, subsequently it requires to be upgraded.

From the aforementioned examples, as well as from number of completed experiments with 'raw data', it is evidently that universal function per diagnostic information is not exist yet (either we could not find it!). Our next step was checking procedure of existing libraries for the task of anomalies detection by means of R-language and Python.

From the whole available libraries relating R-language and Python, the best was Forecast Library of R-language, partially tsoutliers() from the above library (it is the one

which is being used by us now for spikes annihilation). On Fig. 9 you may watch achieved results from present function implementation while data filtration from mentioned sensors.

Form completed by us experiments with various options regarding filtration performance we did consider the best one from the above is the last one, as well as it gives acceptable result reckoning anomalies filtration.

IV. CONCLUSION

For quality analysis of diagnostic information completion being acquired from monitoring systems of any spheres, prior filtration of initial data routine should be prevailed. 'Raw info data' may be composed of behemoth matrix with a lot of anomalies being included. Existence of the abovementioned deviations seriously aggravates further data processing performance aimed on exact diagnosis determination with subsequent solution of geneses tasks and later forecasting phase. Effectiveness resume of monitoring systems with filtration function deficiency is considered as a huge amount of piled information without any kind of worth analysis execution, where data processing operations are being worked out compares to some terminal values (designated limits, norms, etc.) without specific estimation of blackouts expansion trends as well as characteristics assessment of being controlled site. Consequently this way, monitoring systems were being 'hit an iron ceiling', being just the storage of immense amount of data, unfortunately, without essential instruments of safe and reliable indices improvement, which are so crucial for sophisticated technical structures services.

Experiments with a lot massive files of initial diagnostic extract, which were arranged by authors of present article, proved the idea of search option gain of proper filtration methods realization. Let us highlight the following results:

- by means of statistic methods we did handle the filtration of 'raw diagnostic data' within 50% of presented cases (more than 500 diagrams of 'raw diagnostic data';
- via 'machine learning' methods we managed successfully completed the purge of 'raw diagnostic data' about of 60% cases;
- by means of 'time-series' we succeed purification of 'raw diagnostic data' is 70% cases;
- relating filtration from what was available in library with R-language, it was about a 98% cases (as for the rest of instances, there were some negligible errors).

We shall mention, that for this particular task solution, one may apply autoencoders-types of artificial neural network used to learn efficient data coding in an unsupervised manner via application of backpropagation of error [26]. Hence, those methods should be analyzed after all the others, for the reason of complicated adjustment per each particular case either features choice performance for every sensor must be automated.

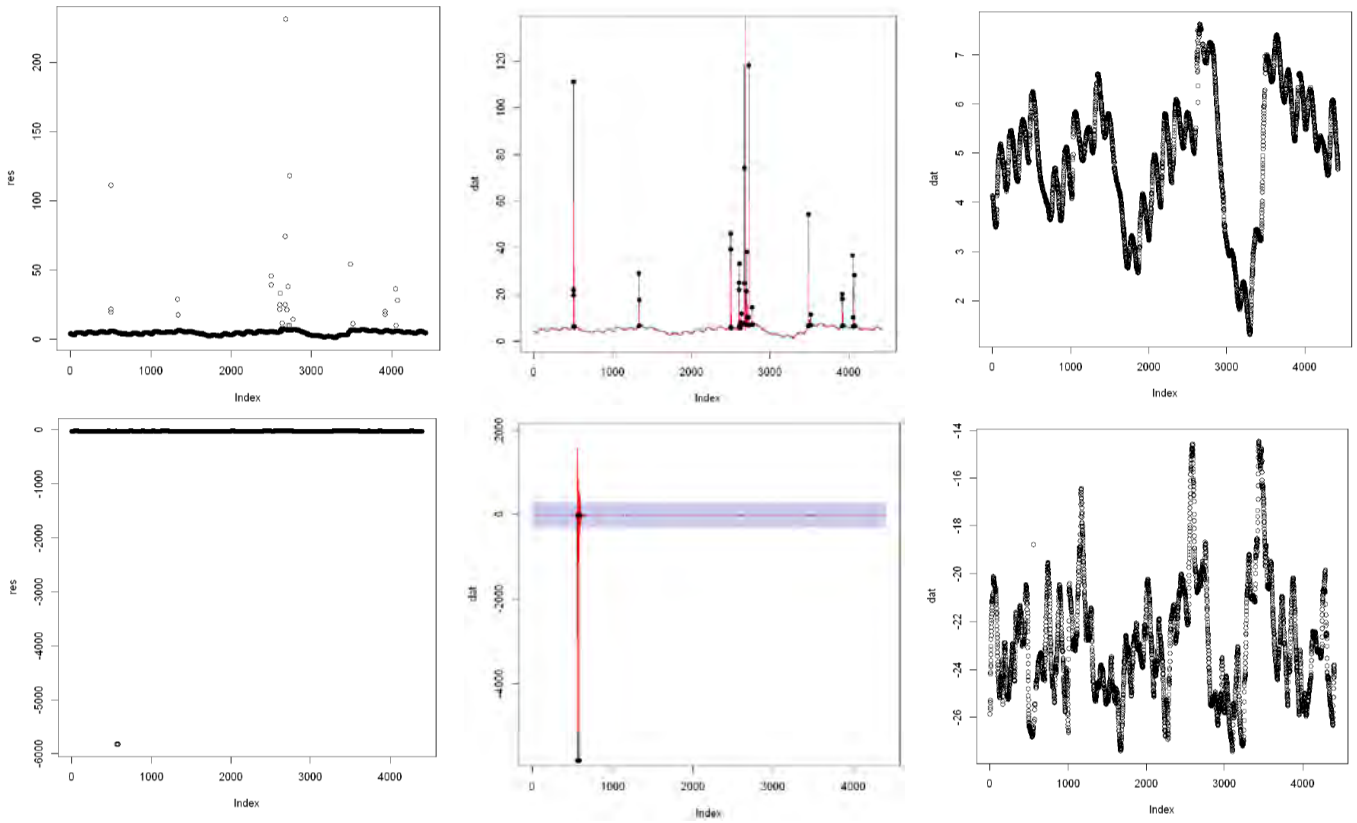


Fig. 8. Filtration data stages by means of R-language form RBG4DP sensor (ETS, upper diagrams) and M6L2DP (ARIMA, lower diagrams): from the left to the right there is an initial diagram, graphic with highlighted anomalies, 'clean diagram'.

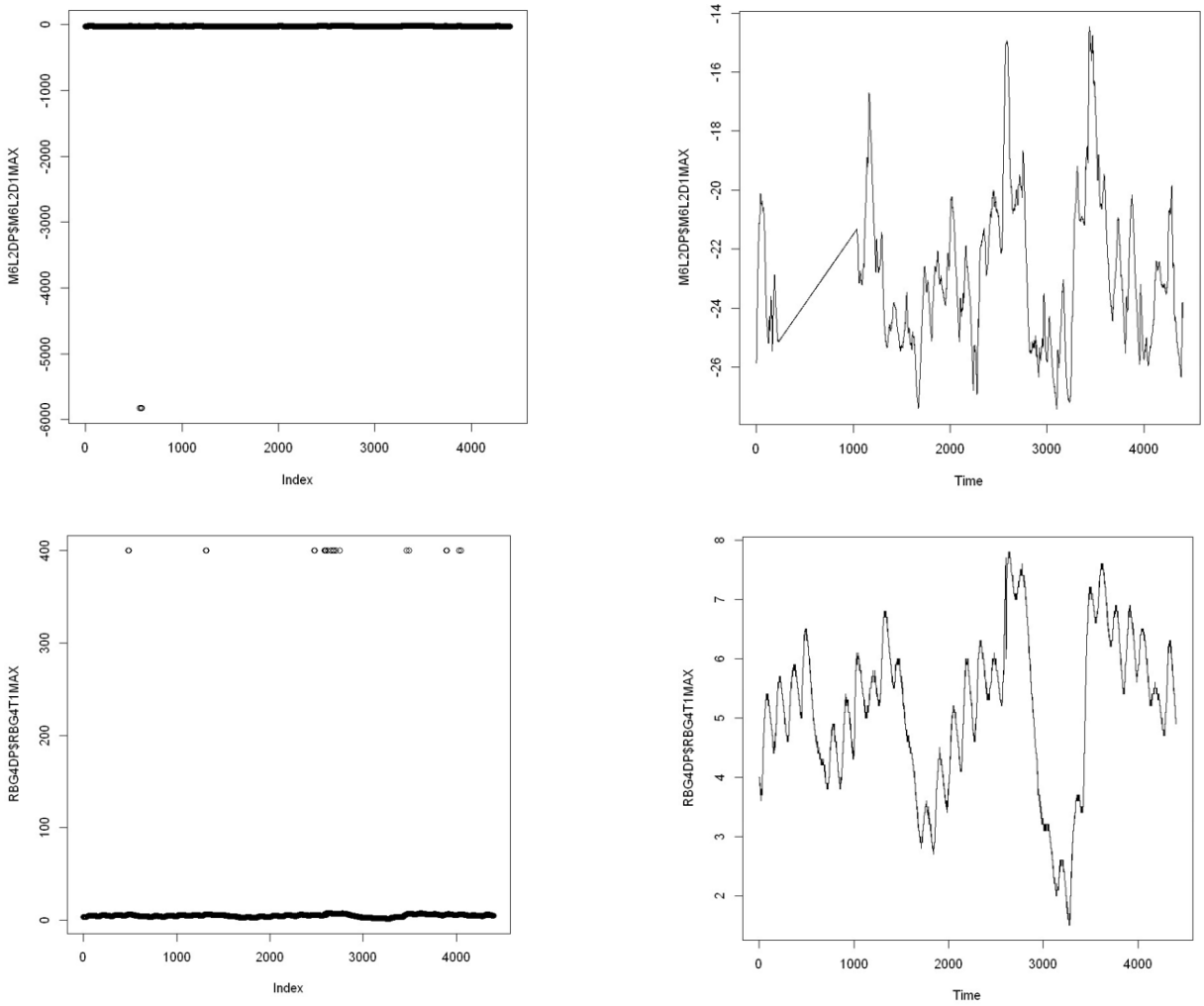


Fig. 9. Filtered data of inserted function in Forecast Library with R- language: upper diagrams – 'raw data' with filtered data from M6L2DP strain gauge; lower diagram – 'raw data' and filtered data from RBG4DP - temperature sensor.

Let us once more bring to notice that within present article we are considering processing performance of already existing diagnostic data (retrospective processing or posterior processing), which is typical for systems of big number of being poled sensors, therefore bringing us a large volumes of information. For processing routine of the above systems in real time mode (for example, those type of monitoring systems are wide spread in transportation branch [27, 28]) other approaches ought to be practical. For instance, preliminary purification of incoming diagnostic information via previous separation of, evidently false measurements, with later restoration of true data, in the event of anomalies availability, from some kind of single sensor. Handling this challenge requires a special data processing architecture arrangement, which allows us not to annihilate useful files prior to entering it to storage slot of diagnostic information.

Monitoring technologies are permanently being improved as relating to components elements base of monitoring systems, as well as regarding processing methods of outcome diagnostic information. We believe, filtration of 'raw diagnostic' files – is the stepping stone reckoning actual 'clean' data derivation with afterward processing operation with a help of AI implementation toward suitable solutions of significant challengers of monitoring performances via the formula 'diagnosis – forecast – remaining life time'.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] A.A. Belyi, E.S. Karapetov, and Yu.I. Efimenko "Structural Health and Geotechnical Monitoring During Transport Objects Construction and Maintenance (Saint-Petersburg Example)", Procedia Engineering, 2017, Vol. 189, pp. 145-151, doi: 10.1016/j.proeng.2017.05.024.

[5] K. Smarsly, and D. Hartmann "Autonomous Monitoring of Masonry Dams Based on Multi-Agent Technology", The 4th Congress on Dams, Struga, Republic of Macedonia, 28-30 September 2017, pp. 1-10.

[6] A. Belyi, G. Osadchy, and K. Dolinskiy "Practical Recommendations for Controlling of Angular Displacements of High-Rise and Large Span Elements of Civil Structures", Proceedings of 16th IEEE East-West Design & Test Symposium (EWDTS`2018), Kazan, Russia, September 14-17, 2018, pp. 176-183, doi: 10.1109/EWDTS.2018.8524743.

[7] A. Belyi, D. Shestovitskii, V. Myachin, and D. Sedykh "Development of Automation Systems at Transport Objects of MegaCity", Proceedings of 17th IEEE East-West Design & Test Symposium (EWDTS`2019), Batumi, Georgia, September 13-16, 2019, pp. 201-206, doi: 10.1109/EWDTS.2019.8884382.

[8] M. Wernet, M. Brunokowski, P. Witt, and T. Meiwald "Digital Tools for Relay Interlocking Diagnostics and Condition Assessment", Signal+Draht, 2019 (111), issue 11, pp. 39-45.

[9] S.V. Mikoni, B.V. Sokolov, and R.M. Yusupov "Qualimetry of Models and Polymodel Complexes" (in Russ), Moscow, the Russian Academy of Sciences (RAS), 314 p.

[10] J.E. Andersen, and A. Vesterinen "Structural Health Monitoring Systems", COWI, 2006, L&S S.r.l. Servizi Grafici,125 p.

[11] T. Asada "Novel Condition Monitoring Techniques Applied to Improve the Dependability of Railway Point Machines", University of Birmingham, UK, Ph. D. thesis, May 2013, 149 p.

[12] W. Jin, Z. Shi, D. Siegel, P. Dersin, C. Douziech, M. Pugnaloni, P. La Cascia, and J. Lee "Development and Evaluation of Health Monitoring Techniques for Railway Point Machines", 2015 IEEE Conference on Prognostics and Health Management (PHM), 22-25 June 2015, Austin, TX, USA, doi: 10.1109/ICPHM.2015.7245016.

[13] T. Böhm "Remaining Useful Life Prediction for Railway Switch Engines Using Artificial Neural Networks and Support Vector Machines", International Journal of Prognostics and Health Management 8 (Special Issue on Railways & Mass Transportation), December 2017, 15 p.

[14] C. Cremona, and J.P. Santos "Structural Health Monitoring as a Big-Data Problem", Structural Engineering International, 2018, vol. 28, issue 3, pp. 243-254, doi: 10.1080/10168664.2018.1461536.

[15] D. Luckey, H. Fritz, D. Legatiuk, K. Dragos, and K. Smarsly "Artificial intelligence techniques for smart city applications", Proceedings of the International ICCCBE and CIB W78 Joint Conference on Computing in Civil and Building Engineering 2020, Sao Paolo, Brazil, 06.02.2020, pp. 1-14.

[16] T. Neumann, D.N. Guzmán, and J.C. Groos "Transparent Failure Diagnostics for Railway Switches Using Bayesian Networks", Signal+Draht, 2019 (111), issue 12, pp. 23-31.

[17] S.A. Sokolov, D.G. Plotnikov, A.A. Grachev, and V.A. Lebedev "Evaluation of Loads Applied on Engineering Structures Based on Structural Health Monitoring", International Review of Mechanical Engineering (IREME), 2020, Vol. 14, No. 2, pp. 146-150.

[18] List of Longest Cable-Stayed Bridge Spans: https://en.wikipedia.org/wiki/List_of_longest_cable-stayed_bridge_spansıo

[19] S.A. Glantz "Primer of Biostatistics", 7th edition, New York, McGraw-Hill Medical, 2012, 312 p.

[20] N. Cristianini, and J. Shawe-Taylor "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000, doi: 10.1017/CBO9780511801389.

[21] Sklearn.svm.OneClassSVM: https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html

[22] Sklearn.svm.IsolationForest: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html.

[23] L. Breiman "Random Forests", Machine Learning, 2001, vol. 45, issue 1, pp. 5-32, doi: 10.1023/A:1010933404324.

[24] R.J. Hyndman "Forecasting: Principles & Practice", University of Western Australia, 23-25 September 2014, 138 p.

[25] J.O. Irwin "On a Criterion for the Rejection of Outlying Observations", Journal of Biometrics, 1925, Vol. 17, No 3/4, pp. 238-250.

[26] C.-Y. Liou, C.-W. Cheng, J.-W. Liou, and D.-R. Liou "Autoencoder for Words", Neurocomputing, 2014, Vol. 139, pp. 84-96, doi:10.1016/j.neucom.2013.09.055.

[27] L. Heidmann "Smart Point Machines: Paving the Way for Predictive Maintenance", Signal+Draht, 2018 (110), issue 9, pp. 70-75.

[28] D.V. Efanov, G.V. Osadchy, V.V. Khóroshev, and D.A. Shestovitskiy "Diagnostics of Audio-Frequency Track Circuits in Continuous Monitoring Systems for Remote Control Devices: Some Aspects", Proceedings of 17th IEEE East-West Design & Test Symposium (EWDTS`2019), Batumi, Georgia, September 13-16, 2019, pp. 162-170, doi: 10.1109/EWDTS.2019.8884416.

Measurement and Compact Modeling of Noise Characteristics in Complementary Junction Field-Effect Transistors

Alexandr M. Pilipenko
Department of Fundamentals of Radio Engineering
Southern Federal University
Taganrog, Russia
ampilipenko@sfedu.ru

Fedor A. Tsvetkov
Department of Fundamentals of Radio Engineering
Southern Federal University
Taganrog, Russia
fa-tsvet@yandex.ru

Nikolay N. Prokopenko
Department «Information Systems and Radio Engineering»
Don State Technical University
Rostov-on-Don, Russia
prokopenko@sssu.ru

***Abstract*—The noise of complementary junction field-effect transistors (JFETs) was measured in the frequency range from 1 Hz to 5 kHz for various static modes using the developed and verified hardware-software system on the base of National Instruments platform. The JFETs were developed by JSC "INTEGRAL" (Minsk, Belarus) using silicon BiJFET-technology on which the manufacture of more than 20 types of analog integrated circuits for various sensors and transducers is based. The Features of the complementary JFETs noise characteristics in various frequency bands of the measurement range are found. Parameter extraction and error estimation are realized for the compact SPICE-model of noise in the JFETs. The results of JFETs parameter extraction show a good correspondence of the measured characteristics to the compact SPICE-model of noise. The relative root-mean-square (RMS) error of compact modeling does not exceed 8%.**

Keywords—low noise sensors interfaces, SPICE-model, junction field-effect transistor, JFET, parameters extraction, thermal noise, flicker noise, power spectral density.

I. INTRODUCTION

Silicon BiJFET-technology is currently actively developed for the manufacture of radiation-hardened integrated circuits (ICs). BiJFET-technology allows to realize complementary junction field effect transistors (JFETs) on the same chip with bipolar transistors [1]. JSC "INTEGRAL" (Minsk, Belarus) has mastered the manufacture of more than 20 types of analog ICs for various purposes on the base of BiJFET-technology [2].

The main application area of ICs with JFETs is the processing of signals from sensors of different physical quantities and transducers on spacecrafts, artificial satellites and robotic systems [3]. JFETs have a minimum level of intrinsic noise in comparison with other types of transistors, so the use of JFETs in the input stages of ICs allows to realize ultra-low-noise operational amplifiers [4].

Measurement of noise in JFETs is a rather complex engineering problem. Firstly, to identify the SPICE-model parameters of noise in JFETs, it is necessary to measure the current noise power spectral density (PSD) at very low frequencies (from 1 Hz), secondly, the values of JFETs current noise PSD can be very small (10^{-22} A^2/Hz or less) [5]. Standard spectrum analyzers are not suitable for measuring JFETs noise, because they have a lower limit of the frequency span equal to 9 kHz and sensitivity about 120 dBm, which corresponds to PSD $W = 10^{-18}$ A^2/Hz at the resistive load $R_s = 24$ kΩ [6]. Thus, to measure JFETs noise characteristics, it is necessary to create specialized highly sensitive hardware-software systems [7].

The goal of this work is to determine the parameters of the compact SPICE-model of noise in complementary JFETs, developed by JSC "INTEGRAL" (Minsk, Belarus) on the base of BiJFET-technology. To achieve this goal the following problems are solved:

- development of the hardware-software system to measure noise characteristics in the complementary JFETs;

- measurement and analysis of the complementary JFETs noise characteristics in the frequency range from 1 Hz to 5 kHz for various static modes;

- parameter extraction and error estimation of the compact SPICE-model of noise in the complementary JFETs according to measured noise characteristics.

II. COMPACT SPICE-MODEL OF NOISE IN JFETS

JFETs noise characteristics in SPICE simulators are described using the compact SPICE-model of drain current fluctuations (current noise), which has the form of current source. PSD of the current noise in SPICE has the form [8]:

$$W_{ID}(f) = \frac{8kTS}{3} + \frac{KF \cdot I_D^{AF}}{f}, \qquad (1)$$

where f is the frequency; S is the differential transconductance of the transistor at the operating point; I_D is the DC component of the drain current at the operating point; KF is the coefficient which defines flicker noise PSD; AF is the exponent which defines the dependence of flicker noise PSD upon the drain current; $k = 1.38 \cdot 10^{-23}$ J·K^{-1} is the Boltzmann constant, T is the absolute temperature in Kelvin (K).

The reported study was supported by the Grant of the Russian Science Foundation according to the research project No. 16-19-00122-P

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

The first part in the equation (1) describes the JFETs channel thermal noise PSD:

$$W_{IDT} = \frac{8kTS}{3}. \quad (2)$$

The second part in the equation (1) describes the JFETs channel flicker noise (1/*f*- noise) PSD:

$$W_{IDF}(f) = \frac{KF \cdot I_D^{AF}}{f}. \quad (3)$$

The channel thermal noise dominates in the frequency range $f \geq 100$ Hz and does not depend on the frequency. The channel flicker noise dominates in the frequency range $f < 100$ Hz and significantly depends on the frequency.

It should be noted that the thermal noise PSD of resistance $R = 1/S$ has the form:

$$W_{IR} = \frac{4kT}{R}. \quad (4)$$

As we can see from equations (2) and (4) the channel thermal noise PSD is 1.5 times smaller than the thermal noise PSD of the resistance $R = 1/S$. This property is explained by the fact that the channel is nonuniform and is not in the thermal equilibrium state in the saturation region.

III. Hardware-Software System For Measurement of Noise in JFETs

The hardware-software system for measurement of noise in JFETs was realized in this work on the base of National Instruments (NI) hardware and LabVIEW software. The block-diagram of setup for measurement of noise in JFETs is shown in Fig. 1.

The measurement setup comprises the following units:

- *LPF1* and *LPF2* are the low-pass filters with the cutoff frequency no more than 1 Hz for suppressing the noise from the DC voltage sources of V_{GS} and V_{DS}, where V_{GS} is the gate-to-source voltage and V_{DS} is the drain-to-source voltage;
- *DA* is the operational amplifier which forms with the resistor R_s a current-voltage converter and maintains the V_{DS} voltage specified level;
- C_b is the blocking capacitor;
- R_s is the low-noise resistor which converts the drain current to the *LNA* input voltage;
- *HPF* is the high-pass filter with the cutoff frequency no more than 1 Hz for suppressing the DC voltage from the *DA* output;
- *LNA* is the low-noise amplifier which provides the specified level of voltage at the *DAQ* input;
- *DAQ* is the data acquisition device which provides analog-to-digital conversion (ADC) of the *LNA* output voltage;
- *PC* is the personal computer which calculates the current noise PSD.

The data acquisition device was realized on the base of NI USB-6251 module, which has the following characteristics [9]:

- ADC resolution $m = 16$ bits;
- maximum sampling frequency $f_D = 1.25$ MHz;
- input voltage range from ± 0.1 V to ± 10 V.

For the correct functioning of the measurement setup the following condition must be executed:

$$I_D R_s + V_{DS} \leq E_0 - 1{,}5 \text{ V},$$

where E_0 is the supply voltage for *DA*.

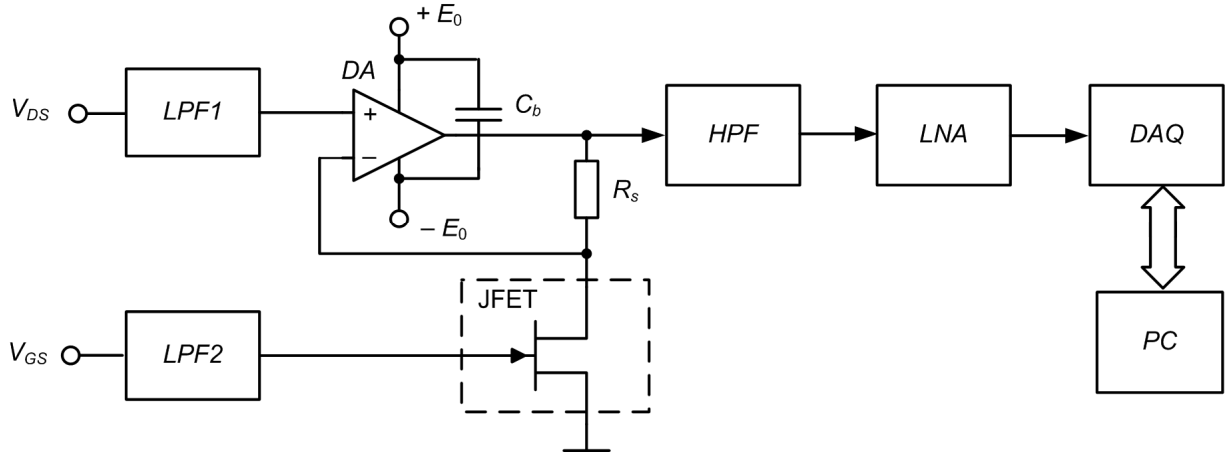


Fig. 1. Block-diagram of setup for measurement of noise in JFETs

The experimental noise characteristic of the transistor is determined on the base of the *LNA* output voltage samples sequence using LabVIEW software. To obtain the experimental noise characteristic the following operations are performed:

1. Fast Fourier transform (FFT) of the *LNA* output voltage samples sequence;

2. Determination of the current noise PSD using the measured root-mean-square (RMS) values of the *LNA* output voltage harmonic components

$$W_{ID,k} = \frac{U_k^2}{\Delta f \, R_s^2 \, K_u^2}, \quad (5)$$

where U_k is the RMS value of the *LNA* output voltage harmonic component at $f = f_k$; Δf is the FFT frequency spacing; K_u is the *LNA* gain.

The *LNA* output voltage is inputted to the *PC* via NI USB-6251 module. Duration of the *LNA* output voltage realization is equal to 1 s, so the FFT frequency spacing is $\Delta f = 1$ Hz. The sampling frequency is chosen equal to 500 kHz, so one realization with duration of 1 s contains 500 000 points. The measured values of the PSD $W_{ID,k}$ are averaged over the set of the input realizations. The number of realizations while measuring the noise characteristic in the frequency range from 1 Hz to 5 kHz was about 1000 for each operating point of the transistor.

IV. Parameter Extraction of Compact SPICE-model of Noise in JFETs

The JFETs noise SPICE-model parameters *KF* and *AF* are determined by the least squares method if the sum of squares of the relative errors of the current noise PSD approximation is minimal:

$$F = \sum_{k=1}^{N} \left[\frac{W_{ID}(f_k) - W_{ID,k}}{W_{ID,k}} \right]^2, \quad (6)$$

where N is the number of measured values of the current noise PSD; $W_{ID,k}$ is the measured values of the current noise PSD; $W_{ID}(f_k)$ is the values of the current noise PSD calculated for $f = f_k$ using model (1).

The minimum of the function (6) was determined on the base of the Levenberg-Marquardt algorithm with accuracy control under variation of initial conditions. The accuracy of modeling was estimated using the RMS error of the SPICE-model:

$$\sigma = \sqrt{\frac{F_{\min}}{N}}, \quad (7)$$

where $F_{\min}$ is the minimum value of the function (6).

Measurement and modeling of the noise characteristics was carried out for the tested complementary JFETs of JSC "INTEGRAL" (n-channel JFET – nJFET and p-channel JFET – pJFET) at room temperature. Fabrication technology of the tested JFETs is described in [10], the measured I-V characteristics are presented in [11] and [12]. Dimensions of the tested transistors are as follows:

$W = 260$ μm, $L = 8$ μm – for nJFET;

$W = 50$ μm, $L = 6$ μm – for pJFET,

where W is the channel width; L is the channel length.

The tested JFETs are intended to be used in ICs of operational amplifiers which can operate under simultaneous influence of extremely low temperatures and radiation [12].

The noise characteristics of the tested JFETs were measured at several operating points of the saturation region. The drain-to-source voltage for each tested transistor was kept constant: $V_{DS} = 5$ V for nJFET and $V_{DS} = -5$ V for pJFET. The gate-to-source voltage varied with the step of 0.25 V within the following limits: $V_{GS} = -0.75 \ldots 0$ V for nJFET and $V_{GS} = 0 \ldots 0.75$ V for pJFET. The values of the JFETs drain current and the differential transconductance corresponding to the selected values of V_{DS} and V_{GS} are shown in Fig. 2. It should be noted that the drain current in the selected voltage range varies by a factor of about 8 times for nJFET and about 5 times for pJFET.

Fig. 3 presents the results of measurement and modeling of the current noise PSD of the tested JFETs in various operating points at room temperature $T = 27$ °C. Similar measurement results were obtained for three pairs of complementary JFETs of JSC "INTEGRAL". To reduce the influence of external noise on the measurement results and to increase the accuracy of the measurement setup, accumulators should be used as the voltage sources of V_{GS} and V_{DS}, it is recommended to use metal-film resistors and high-quality capacitors to build filters, the measurement setup should be shielded by placing in a grounded metal case.

As we can see from Fig. 3 the measured current noise PSD of the tested JFETs at $f > 1$ kHz correspond to the thermal noise model (2) and have the following values:

$W_{ID,k} \approx (3{,}5 \ldots 12) \cdot 10^{-24}$ A^2/Hz – for nJFET;

$W_{ID,k} \approx (1.5 \ldots 3.5) \cdot 10^{-24}$ A^2/Hz – for pJFET.

The levels of $W_{ID,k}$ at $f > 1$ kHz almost coincide with the thermal noise PSD values of W_{IDT} calculated by the formula (2) at the chosen operating points (the relative deviation of the measured $W_{ID,k}$ values from the calculated W_{IDT} values does not exceed 6 %). This fact confirms the reliability of the measurement results. In addition, to verify the measurement setup, the current noise of the resistor $R_s = 24$ kΩ was determined. The measured value of the current noise PSD for the resistor R_s in the frequency range from 10 Hz and higher differed from the calculated value of W_{IR} by no more than 5%.

Fig. 3 shows that the frequency dependences of the measured current noise PSD for the tested JFETs are described by the following laws:

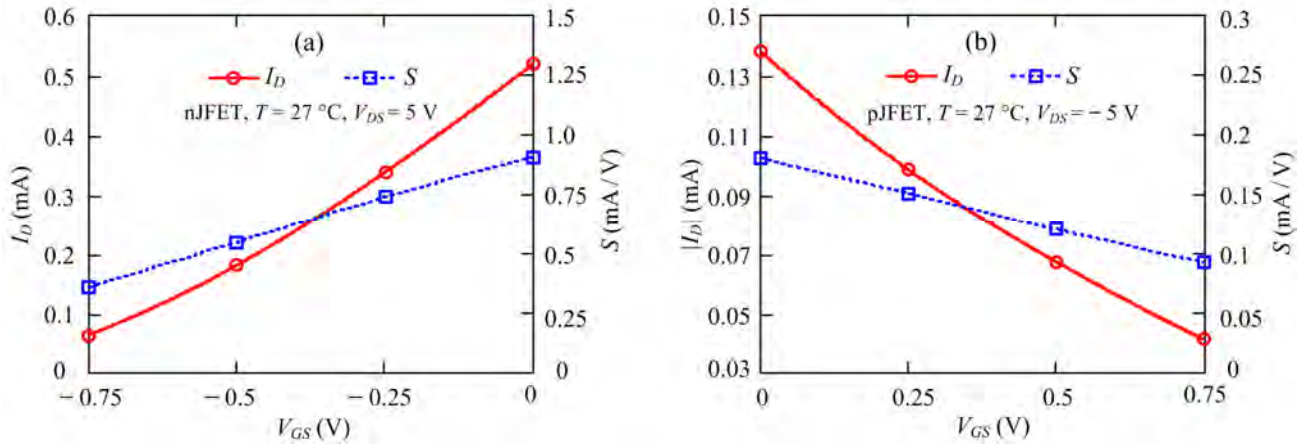


Fig. 2. Static I-V characteristics and differential S-V characteristics of nJFET (a) and pJFET (b)

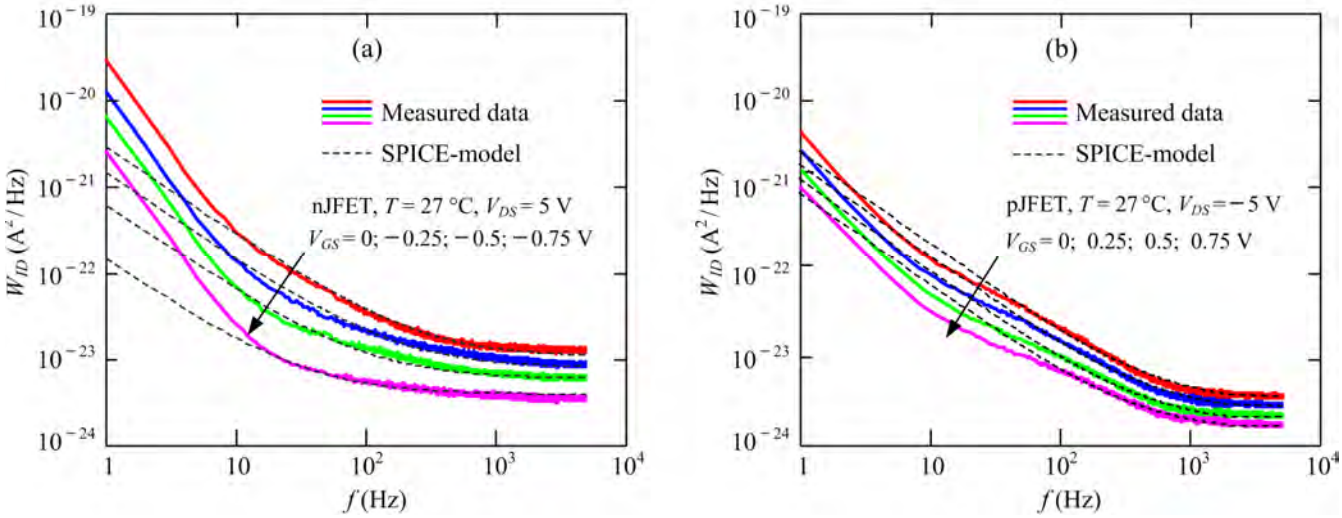


Fig. 3. Measured power spectral density of noise in nJFET (a) and pJFET (b)

$1/f^2$ in the frequency range from 1 to 10 Hz and according to the SPICE model (1) in the frequency range from 10 Hz and higher – for nJFET;

$1/f^{1.5}$ in the frequency range from 1 to 10 Hz, $1/f^{0.75}$ in the frequency range from 10 to 100 Гц and according to the SPICE model (1) in the frequency range from 100 Hz and higher – for pJFET.

The above-described properties of JFETs are explained by their manufacture technology and the mobility of the majority charge carriers in the channel. It should be noted that the commercial JFETs described in [5] have similar properties.

The parameter extraction of the SPICE-model of noise in JFETs (*KF* and *AF*) was carried out by minimizing the function (6). To extract the *KF* and *AF* parameters, only $W_{ID,k}$ values which lie in the frequency range where the measured PSD corresponds to the SPICE-model, were used: f_k = 100 Hz ... 5 kHz for nJFET and f_k = 100 Hz ... 5 kHz for pJFET (see Fig. 3). As a result of minimizing the function (6), the parameters of the SPICE-model and the RMS errors were obtained:

$KF = 1.12 \cdot 10^{-16}$, $AF = 1.408$, σ = 7.7 % – for nJFET;

$KF = 2.77 \cdot 10^{-18}$, $AF = 0.845$, σ = 6.5 % – for pJFET.

V. Conclusions

The hardware-software system, which allows to measure the intrinsic noise of complementary JFETs in the frequency range from 1 Hz to 5 kHz has been developed and verified on the base of the National Instruments platform. The hardware-software system has a large algorithmic flexibility during measurements. Recommendations have been formulated to reduce the influence of external noise on the measurement results and to improve the accuracy of the measurement setup.

The frequency dependences of the current noise PSD have been measured for the tested complementary JFETs of JSC "INTEGRAL" (Minsk, Belarus). It has been shown that the thermal noise PSD level in complementary JFETs is approximately 10^{-24} A²/Hz which is two orders of magnitude lower than the thermal noise of commercial JFETs described in [5].

The features of the nJFET and pJFET noise characteristics in the frequency bands 1 … 10 Hz, 10 … 100 Hz and 100 Hz … 5 kHz have been identified. The measured noise characteristics of the complementary JFETs over the most part of the frequency range 1 Hz ... 5 kHz correspond to the compact SPICE-model of current noise in JFETs. The relative RMS error of compact modeling does not exceed 8%.

REFERENCES

[1] O. V. Dvornikov, V. L. Dziatlau, V. A. Tchekhovski, N. N. Prokopenko and A. V. Bugakova, "BiJFet Array Chip MH2XA030 — a Design Tool for Radiation-Hardened and Cryogenic Analog Integrated Circuits," 2018 IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), St. Petersburg, 2018, pp. 13-17, doi: 10.1109/EExPolytech.2018.8564415.

[2] A. V. Bugakova, D. Y. Denisenko, N. N. Prokopenko and A. E. Titov, "Basic Functional CJFet Op-Amp Nodes for Operation at Low Temperatures and at Conditions of Penetrating Radiation," 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), Varna, Bulgaria, 2019, pp. 1-4, doi: 10.1109/ELMA.2019.8771657.

[3] N. Makris, M. Bucher, F. Jazaeri and J. Sallese, "CJM: A Compact Model for Double-Gate Junction FETs," IEEE Journal of the Electron Devices Society, vol. 7, pp. 1191-1199, 2019, doi: 10.1109/JEDS.2019.2944817.

[4] S. S. Li, Semiconductor Physical Electronics, 2nd ed., Springer, 2006.

[5] N. Makris, L. Chevas and M. Bucher, "Compact Modeling of Low Frequency Noise and Thermal Noise in Junction Field Effect Transistors," ESSDERC 2019 - 49th European Solid-State Device Research Conference (ESSDERC), doi: 10.1109/ESSDERC.2019.8901775.

[6] Spectrum analyzer GW INSTEK, [online] Available: http://vebion.ru/catalog/kontrolno_izmeritelnye_pribory/analizatory_spektra/analizator_spektra_gw_instek/

[7] K. Kellogg, L. Dunleavy, S. Skidmore, H. Morales and C. White, "Bridging the Gap in noise spectral density measurements derived from flicker and noise figure measurement systems," 2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON), Cocoa Beach, FL, 2017, pp. 1-4, doi: 10.1109/WAMICON.2017.7930268.

[8] P. Horowitz, W. Hill, The Art Of Electronic, 3rd ed., Cambridge University Press, 2015.

[9] NI 6251 Device Specifications, [online] Available: https://www.ni.com/pdf/manuals/375213c.pdf

[10] I.Y. Lovshenko, V.T. Khanko and V.R. Stempitsky, "Radiation influence on electrical characteristics of complementary junction field-effect transistors exploited at low temperatures," Materials Physics & Mechanics, vol. 39. no. 1, pp. 92–101, 2018, doi: 10.18720/MPM.3912018_15.

[11] A. M. Pilipenko, V. N. Biryukov and N. N. Prokopenko, "A Template Model of Junction Field-Effect Transistors for a Wide Temperature Range," 2019 IEEE East-West Design & Test Symposium (EWDTS), Batumi, Georgia, 2019, pp. 1-4, doi: 10.1109/EWDTS.2019.8884411.

[12] V. N. Biryukov, A. M. Pilipenko, N. N. Prokopenko, O. V. Dvornikov, "Template model of complementary field-effect transistors with a control pn junction," Zhurnal Radioelektroniki – Journal of Radio Electronics, no. 8, 2019. doi: 10.30898/1684-1719.2019.8.5 [in Russian].

Hardware implementation of timed logical control FSM

Maryna Miroschnyk
professor, USURT
Kharkiv, Ukraine
marinagmiro@gmail.com

Alexander Shkil
associate professor, KhNURE
Kharkiv, Ukraine
oleksandr.shkil@nure.ua

Dariia Rakhlis
associate professor, KhNURE
Kharkiv, Ukraine
dariia.rakhlis@nure.ua

Elvira Kulak
associate professor, KhNURE
Kharkiv, Ukraine
elvira.kulak@nure.ua

Inna Filippenko
associate professor, KhNURE
Kharkiv, Ukraine
inna.filippenko@nure.ua

Mykyta Malakhov
student, KhNURE
Kharkiv, Ukraine
mykyta.malakhov@nure.ua

Abstract – Methods of hardware implementation of event-driven timed control FSM were considered in the article. Classification of timed control FSM into active and passive by the method of processing input signals, and into Moore and Mealy models by the method of generating output signals were given. Timing parameters are implemented by the counter of FSM' clock cycles. Timed FSM models are presented in the VHDL hardware description language in the form of automata pattern. Behavioral simulation of proposed models, synthesis and implementation in FPGA, as well as simulation after implementation using CAD Xilinx ISE 14.7 were performed.

Keywords − timed finite state machine, Mealy model, Moore model, state diagram, hardware description language, automata patterns, waveform, CAD, XILINX ISE.

I. Introduction

Among the whole set of control systems, a significant part are logical control systems, in which control signals take logical zero or one depending on the boundary values of physical quantities that determine these parameters. For the technical implementation of these systems, the most suitable model is a structural finite state machine (FSM), and the state diagram is a visual representation of the functioning algorithm. A distinctive feature of finite state machines in logical control systems is the presence among input values not only signals of the control object, but also external, to the controlled system, signals (external events), which provide the interaction of the logical control system with the external environment [1].

Most real logic control systems are real-time systems and for their implementation it is customary to use the model of a timed FSM and for a visual representation – the temporal state diagram. During automated synthesis of timed FSM using hardware description languages for the correct synthesis of the FSM circuit taking into account timing parameters, as a rule, automata' programming patterns in HDL are used. For the VHDL language, this is a special structure of the VHDL model, in which the transition and output functions are allocated to separate processes (process), the new state is assigned in a special process related to synchronization, and delays are realized by means of the counter of the FSM' clock cycles. In addition, during processing external events in real-time systems, it is necessary to take into account the time period during which external events can change the algorithm of the control FSM. The aim of this work is to develop a single pattern in the VHDL language for describing different types of timed control FSM in real-time logical control systems in the style of automata-based programming, which are implemented on the FPGA hardware platform.

II. Formal Statement of the Problem

Let the generalized model of a structural timed control FSM was given as $Y(t) = g(X(t), Z(t), T)$, $Z(t+1) = f(X(t), Z(t), T)$, where X – set of input values, Z – set of internal variables, which coding style is determined FSM states, and Y – set of output values, t – FSM time, which is determined in clock cycles, g – is the outputs function of the structural FSM, f – is the transitions function of the structural FSM. $T = \{t_c, t_o, t_d\}$ – are timing parameters of the FSM, namely: timing constraints t_c, (input) timeouts t_o, and output delays t_d.

Visual formalism of the designed timed FSM specification is temporal state diagram and timing diagram (waveform). The temporal state diagram defines a list of input variables, output variables and FSM states, as well as delays and timing constraints, i.e. is a complete mathematical model of a timed FSM. Waveform reflects the functioning law of the controlled device in FSM's time.

During the hardware implementation of the timed control FSM the problem of external signals and events processing, that appeared at random moments of the time, occurs. To solve the task it is necessary to introduce a classification of timed FSMs models in accordance with input signals (events) processing method and the method of output signals generation. For each type of the FSM model, it is necessary to develop a pattern of the synthesized subset of the VHDL-model for this FSM in the style of automata-based programming, and confirm the correctness of the proposed model by simulation, synthesize and implementation based on FPGA with CAD tool XILINX ISE.

III. Literature Review

Models of finite state machines that are implemented in both hardware and software are widely used in real-time logical control systems.

In [1], the role and place of control FSMs in logical control systems is defined. The technique of designing automatic logical control systems based on real time and processing of external events is given. The classification of external events and methods of their processing are given.

A timed FSM, as a way to implement a control algorithm in real-time systems, was introduced in [2]. The state diagram of the FSM is supplemented by a finite set of timers that take real values. The nodes of the diagram are called positions, and edges are called transitions. Each timer is reset to zero at the time of transition and increases its value with each FSM clock cycle. Each transition is

978-1-7281-1003-5/19/$31.00 ©2020 IEEE

associated with a clock constraint, which means that this transition can only be carried out if the current timer values satisfy this restriction. Each position has a timer constraint called an invariant; the system can be in this position only as long as its invariant is satisfied.

The theory of timed FSM was further developed in works related to testing real-time hardware systems. In [3], timed FSM models (TFSM) that take into account timeouts in states and delays of output signals with respect to the implementation of state transition are considered. At the same time, it is taken into account that if no input signal is received during the timeout, then the FSM goes to the next state. Methods for constructing tests for the composition of TFSM are considered. Considered models were used during time parameters testing of client-server banking systems. In [4], questions of constructing tests for composition of timed FSMs with timed guards and output delays were considered. In this paper how a test suite with the guaranteed fault coverage can be reduced for a system of interacting TFSMs when only some components can be faulty is discussed. Given a component TFSM, a corresponding test is derived for the composition of TFSMs under the assumption that all other components are fault-free. In [5], methods for minimizing models of timed FSMs for systems with timeouts and timing constraints were proposed. Also special cases of using TFSM models only with timeouts or only with timing constraints were considered. Minimization questions of not only states number, but also timing characteristics are considered. In [6], a generalized model of a timed FSM with a single clock that includes timed guards, timeouts and output delays, is considered. Also was derived a procedure to build a minimal form for deterministic TFSMs. That reduces the number of states, the number of transitions and the timeout values at each state, and it is unique up to isomorphism for non-initialized TFSMs.

In [7], the influence of state coding methods in structural models of finite state machines on hardware costs and speed was considered. An approach that allows to use FPGA input and output buffer triggers as memory elements of a finite state machine is proposed. For this purpose, a new classification of structural models of finite state machines has been proposed. In [8] the new formalization of discrete event simulation on the basis of abstract finite state machines, which are transition systems with highly expressive state, has proposed. An operational (transition system) semantics for the two most basic forms of Discrete Event Simulation (DES): event-based simulation (without objects) and object-event simulation is defined. An event-based simulation (ES) model is a triple {SV, ET, R}, where SV is a set of state variable declarations defining the structure of possible system' states; ET is a set of event type definitions; R is a set of event rules expressed in terms of SV and ET. The classification of event type on internal and external, as well as rules for their processing, was introduced. It is shown that a set of event rules correspond to FSM transitions function.

In [9], a new classification system for finite state machine models in accordance with the implementation of transitions and output signals was introduced. All machines are divided into three categories. In first FSM category (regular FSM), transitions depend only on input signals, and values of output signals depend only on states. In second FSM category (timed FSM), transitions depend on the input signals and the time of their appearance, and values of the output signals depend only on the states. In third FSM category (recursive FSM), transitions depend on the input signals and the time of their appearance, and values of the output signals depend on the current state and previous state directly, i.e. for output signals of state a_i, the function $y_i = y_i + y_i$ is realized, where a_i is the previous state of the FSM. For these categories of FSMs, different patterns of HDL models in the VHDL and Verilog languages are presented, as well as results of their simulation in the ModelSim system (from Mentor Graphics) and synthesis using Xilinx ISE.

In computer-aided design of control devices in logical control systems on the FPGA technology platform, models in hardware description languages are used. Authors of this work in [10] proposed an approach to constructing VHDL-models of timed Moore FSM which take into account the influence of external events.

IV. Classification of timed FSMs

For the hardware implementation of logical control devices, a structural FSM model is used. In structural models, the input alphabet X of an abstract FSM is converted to a set of input values, the alphabet of states A is converted to a set of internal variables Z, and the output alphabet Y – to a set of output values. Moreover, all three sets are finite. In real-time systems, a structural FSM is represented by a model of a timed FSM.

The state of the structural control FSM in logical control systems is characterized by a set of output control signals that appear at certain points of time and have certain duration. Transitions from state to state are determined by the type of input signals and the time of their appearance. Based on the method of output signals generation, control FSMs are classified into the Moore and Mealy model, and based on the method of input signals procession – into active and passive.

The transition from state to state in the timed Moore FSM is carried out "instantly", and output signals in the state are formed taking into account delays of their appearance. At zero delay output signals appear at the moment when the FSM gets into the state and don't change during the period when FSM will stay in this state.

The appearance time of output signals in the timed Mealy FSM is determined by the appearance time of input signals (taking into account delays). The lifetime of output signals of the Mealy FSM is determined by the duration of the transition to a new state and at the moment of entering a new state, output signals of this state are reset.

If the FSM simultaneously has output signals, which are typical for Moore and Mealy models, then such FSM is called combined (C-FSM) [11].

An active timed FSM operates depending on the value of the input signal at a certain point of the time (not on a change of the input signal, i.e., the input event). Changing the input signal (input events) does not directly initiate changes in the state of the FSM. The FSM polls input signals at time instants determined by the algorithm of its operation and, thus, implements the transition function. The transition time to a new state of the active FSM is fixed (determined by various delays during different transitions), i.e. static.

The operation of the passive FSM is determined by the input event (the FSM respond not to values of the input signal at a certain point of time, but to an event in a certain period of time). A change in the input signal (event) directly initiates the transition of the FSM to a new state (implements the transition function) and determines the moment output signal appears (implements the output function). From this point of view, a passive FSM can be defined as event-driven FSM. The transition time to a new state of the passive FSM is not fixed, i.e. dynamic. It depends on the time of the input event occurrence, which is, essentially, a random event, during the event waiting period.

Input events can be classified as initiating and interrupting (emergency). Initiating events trigger the transition of the FSM to a new state and the issuance of corresponding output signals. An interrupting event interrupts the transition of the Mealy FSM or interrupts the delay in the state of the Moore FSM. In other words, an interrupt event changes the value of the output signal to its "standard" end. But at the same time, some initiating events can be considered as input events, for example, turning on the power or starting some process in a controlled system.

If a part of input signals of the structural FSM is considering as input actions (which are interrogated), and a part of input signals is considered as events, then such FSM is usually called mixed. When choosing the type of control FSM in logical control systems, both of these parameters should be taken into account.

Any active FSM where there is an interrupting event, in fact, is mixed, and in addition, a certain input signal can be simultaneously considered as an input action or event, which at the same time doesn't violate the above classification.

V. HDL-MODELS OF TIMED MOORE FSM

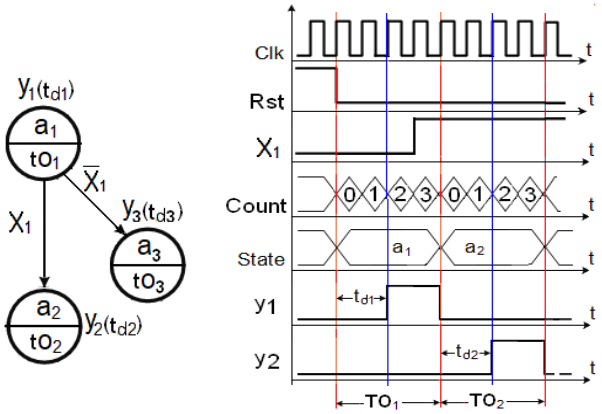
If we review FSM according to the classification by output signals, then Moore FSMs, as a rule, are considered as active or mixed. Such models make up the majority of control FSMs in logical control systems, because it is simple enough to establish a correspondence between technical states of the controlled object and the set of control signals in states of the control Moore FSM. Passive Moore FSMs are rarely considered.

The timeout for the active Moore FSM corresponds to the delay of staying in a certain technical condition. The active Moore FSM reads (polls) the input signal at the time of the active edge of the first clock after the timeout to_i of the current state ends and "instantly" goes to a new state. Output signals of the Moore FSM are related to the current state of the FSM. For each output signal y_j, the delay of its appearance (change) from the moment of transition to the current state is determined. After the timeout, the output signal can be reset to 0 or continued, depending on the algorithm features of the Moore FSM operation. To implement the Moore FSM model with a single time variable in the hardware description language, an additional counter count is used [12].

The timeout to_i is implemented by a multiple transition from state to the same state, while the number of transitions is determined by the number of FSM clock cycles. The value of the counter is compared with (to_i-1), since during the transition to the state a_i, the FSM is staying there for one clock cycle, and therefore to make the timeout exactly equal to to_i clock cycles, (to_i-1) clock cycles are necessary as well. In the temporal state diagram, state timeouts are indicated inside the circles of state symbols and are implemented using loops, conditions for which are value checking of the FSM clock cycles counter. But these loops on the diagram, same as the Clk signal, are not marked.

Delays of output signals in the current i-th state td_{ij} are determined in FSM clock cycles from the moment the FSM goes to the corresponding state for each output signal y_j and are realized by analyzing the counter values, which determines the delay of this signal in FSM clock cycles. On the temporal state diagram, the output delay is indicated in parentheses next to each output signal, and in the VHDL code it is implemented by the conditional signal assignment statement outside the process. The model of the active Moore FSM with $t_o = 1$ and $t_d = 0$ coincides with the model of the traditional microprogram FSM.

Figure 1 shows a fragment of the temporal state diagram of the timed Moore FSM (a), a timing diagram of its functioning (b), and a VHDL code of processes for synchronization, for assignment of new state and new value of the counter (c).



(a) (b)

```
process (Clk, Rst)
  begin
    if Rst = '1' then   state <= a1;
       count <= (others => '0' );
    elsif rising_edge(Clk) then
       state <= next_state; count <= next_count;
    end if;
end process;
process (state, count)
  begin
  next_count <= (others => '0');
  case state is
    when a1 =>
      if count < TO1 - 1 then
        next_state <=a1;
        next_count <= count + 1;
      elsif x1 = '1' then next_state <= a2;
      else  next_state <= a3;
      end if;
  end case;
end process;
y1 <= '1'  when ( state = a1 and count >= td1 ) else '0';
y2 <= '1'  when ( state = a2 and count >= td2 ) else '0';
y3 <= '1'  when ( state = a3 and count >= td3 ) else '0';
```

(c)

Fig.1. Specification and model of the timed Moore FSM

VI. HDL-MODELS OF ACTIVE TIMED MEALY FSMS

Let's consider different variants of timed Mealy FSM models, which can be both active and passive models depending on the class of tasks being solved.

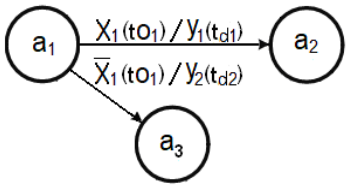
An active Mealy FSM reads (polls) the input signal at the time of the active edge of the first synchronization pulse after FSM goes to the current state. This initiates the beginning of the transition to the next state (a_i –> a_j), the duration of which is determined by the timeout to_i. Until the timeout expires, the FSM stays in the current state. The output signal y_j at the transition appears with a delay td_{ij} and continued until the end of the timeout. After the timeout, the output signal y_j is reset and the FSM goes into a new state. The timeout and the output delay are realized through the analysis of the counter count of the FSM clock cycles.

This model has two features. Firstly, the change in the output signal does not depend on the moment of appearance of the input signal, but is "tied" to the nearest working edge of the clock. This limitation is associated with the constructing feature of VHDL models of timed FSM based on the analysis of the count signal value.

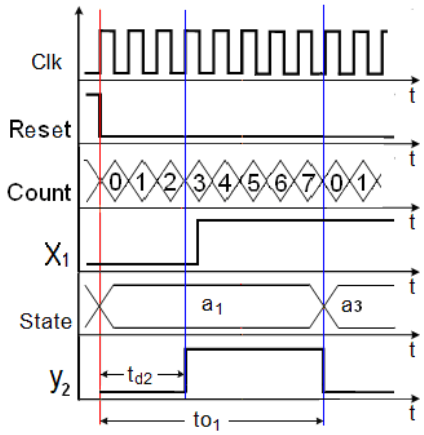
Secondly, the active Mealy FSM is essentially mixed, because input signals are considered as interrogated input events, and in addition to this, interrupting events that interrupt the transition and, accordingly, the output signal, can be used in such model.

The model of the active Mealy FSM is quite close to the traditional (microprogram) FSM, and for $t_o = 1$ and $t_d = 0$ it coincides with it.

Figure 2 shows the specification of the active timed Mealy FSM in the form of the temporal state diagram (a) and timing diagram of its functioning (b), as well as the VHDL code of processes for synchronization, for assignment of a new state and new value of the counter (c).



(a)



(b)

```
-- register input signals and events
process (Clk, Reset)
   begin
     if Reset = '1' then   x1_stored <= '0';
     elsif rising_edge(Clk) then
      if state /= next_state then
        x1_stored <= x1;                          -- x1='0'
      end if;
     end if;
end process;
combinational logic for next_state and Y
process (state, x1_stored, count)
   begin
    y1 <= '0';    y2 <= '0';
    case State is
       when a1 =>
      if x1_stored = '1' then          --transition a1 –> a2
        if count < TO1 - 1 then next_state <= a1;
        else next_state <= a2;
        end if;
        if count >= Td1 then y1 <= '1';
        end if;
      else                                -- transition a1 –> a3
        if count < TO1 - 1 then next_state <= a1;
        else next_state <= a3;
        end if;
        if count >= Td2 then y2 <= '1';
        end if;
      end if;
. . .
    end case;
end process;
```

(c)

Fig. 2. Specification and model of the active Mealy FSM

To implement the model of the active timed Mealy FSM in hardware description language VHDL, it is proposed to allocate for each input signal a trigger signal: signal x_stored. Triggers for input signals x_stored remember the value of the input signal at the time of the first edge of the clock signal when entering the current state. It should be noted that during the synthesis of the circuit, each such signal will correspond to a synchronous trigger.

Thus, the process that forms next_state and y, using the FSM state (state) and values of triggers x_stored, can uniquely determine the current transition. In accordance with the current transition, the timeout, generated output signals, as well as the corresponding output delays, are determined.

Figure 2 shows the implementation of the transition a_1–>a_3 with the output signal y1 (for X = 0 in the first clock cycle Clk) and the value of the trigger x_stored in this case will be '0'. The value of the trigger x_stored = '1' (for X = 1 in the first clock cycle Clk) determines the initialization of the transition a_1 –> a_2 and the output of the signal y_2.

VII. HDL-MODELS OF MIXED MEALY FSM

If, in the active control Mealy FSM model in addition to polled input signals, interrupt external events are used, then this model is usually called mixed.

It is not possible to represent a timed Mealy FSM with interrupting event within the framework of a traditional state diagram. Timing parameters, output signals and event processing are implemented during transitions between

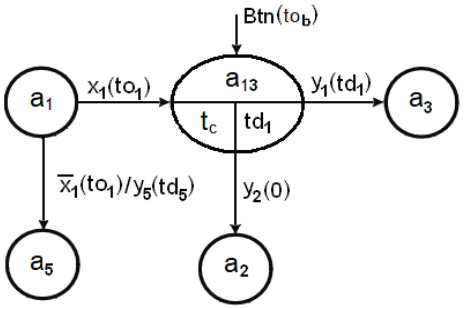
states; and transitions in the state diagram correspond to arcs. Interrupt events processing assumes that for transition e_j must be set the transition e_m, which is initiated as a reaction to the interrupt event. But the traditional state diagram does not assume the presence of an arc corresponding to the transition e_m, since arcs of the state diagram do not implement branching.

It is proposed to use a modification of the state diagram for a timed Mealy FSM with the interrupting event. The state diagram of the timed Mealy FSM – is a bipartite graph with nodes of two types. Nodes of the first type correspond to FSM states and are represented as circles. Nodes of the second type are placed on arcs outgoing from nodes of the first type. A node of the second type implements the processing of interrupting events and timing parameters associated with the corresponding transition, and is represented as an ellipse. For the node of the second type, the FSM state from which the transition is performed, timing constraints, timeout, and output delay for this transition are indicated. If for this transition $tc_{ij} = 0$ or Btn = ∅ (there is no interrupting event), then nodes of the second type can be omitted. If not even one output signal corresponds to the transition under consideration, then for a given node td_{ij} = ∅ (Fig. 3). To implement the model of the active Mealy timed FSM with the interrupting event Btn, a signal-trigger Btn_stored is also introduced for this event.

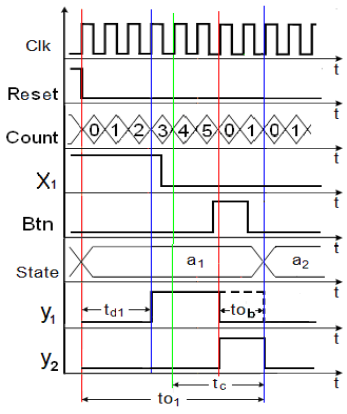
The fact of the interrupting event presence in the interval of timing constraints is determined by values analysis of FSM clock counter (TC0 – 1) <= count < TC1 and is fixed by setting the trigger signal Btn_meet_constraints to '1'. The trigger for interrupting events Btn_stored is set to '1' if the first interrupting event occurs that satisfies timing constraints within the timeout of this state and is reset to '0' during transition between states.

Thus, the process that forms next_state and Y, using the FSM state (state) and values of triggers x_stored and Btn_stored, can uniquely determine the current transition. In accordance with the current transition, the timeout, generated output signals, as well as the corresponding output delays are determined.

Figure 3 shows the specification of the active Mealy FSM with interrupting event Btn in the form of the temporal state diagram (a) and timing diagram of its functioning (b). Figure 3 shows the presence of the interrupting event Btn at the transition a_1–>a_3 in the interval of timing constraints tc and the implementation of a new transition a_1–>a_2 with timeout to_b. In this case, the signal y_2 has a zero output delay.



(a)



(b)

Fig. 3. Specification and model of the active Mealy FSM with the interrupting event Btn

Figure 4 shows the VHDL code of processes for fixing an interrupting event, its falling into the interval of the time limits, and also the assignment of the new state and value of output signals.

```
-- register input signals and events
process (Clk, Reset)
  begin
    if Reset = '1' then
          x1_stored <= '0';
          Btn_stored <= '0';
    elsif rising_edge(Clk)
    then
 if state /= next_state
then      x1_stored <= x1;
Btn_stored <= '0';
 elsif
Btn_meet_constraints =
'1'
   then    Btn_stored <=
'1';
    end if;
    end if;
end process;

-- combination logic for
Btn_meet_constraints
process (state,
x1_stored, Btn, count)
  begin
Btn_meet_constraints <=
'0';
   if Btn = '1' then
      if (state = a1 and
x1_stored = '1' and count
>= TC0 - 1 and count <
TC1) then
Btn_meet_constraints <=
'1';
    end if;
 end if;
end process;
```

```
--combinational logic for
next_state and Y
process (state, x1_stored,
Btn_stored, count)
begin
   y1<='0'; y2<= '0'; y5<= '0';
case State is
when a1 => y2 <= '1';
  if x1_stored = '1' then
    if Btn_stored = '1' then
      if count < TOBtn – 1
      then
            next_state <= a1;
       else  next_state <= a2;
       end if;
     else
       if count < TO1 - 1 then
            next_state <= a1;
       else  next_state <= a3;
      end if;
      if count >= Td1 then
          y1 <= '1';
      end if;
    end if;
   else
       if count < TO1 - 1 then
            next_state <= a1;
      else  next_state <= a5;
       end if;
       if count >= Td5 then
          y5 <= '1';
       end if;
  end if;
. . .
    end case;
end process;
```

Fig. 4. VHDL-model fragments of the active Mealy FSM with the interrupting event Btn

VIII. Experiments and Results

To confirm the reliability of obtained results, behavioral simulation, synthesis and implementation of VHDL models'

fragments for different types of FSMs (fig. 1, 2, 3) using CAD tool Xilinx ISE 14.7 were performed. Figure 5 shows simulation results after implementation of the timed Moore FSM, and figures 6, 7 show simulation results of the active Mealy FSM and the active Mealy FSM with the interrupting event respectively. The simulation results in the form of timing diagrams completely coincided with the specification.

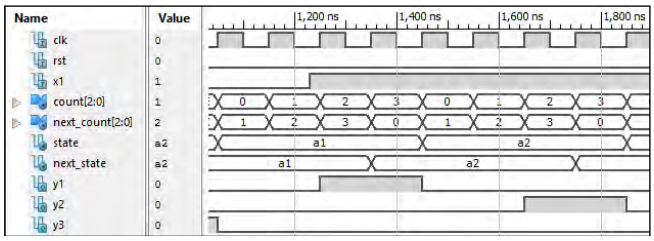


Fig. 5. Simulation waveform of the Moore FSM operation

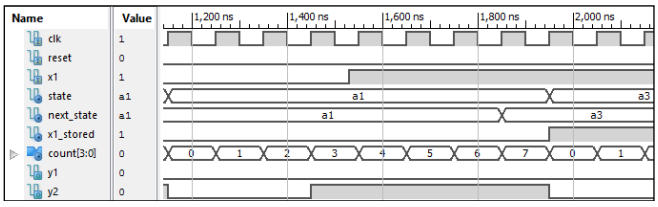


Fig. 6. Simulation waveform of the active Mealy FSM operation

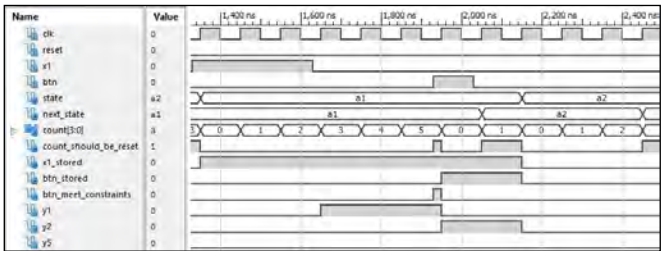


Fig. 7. Simulation waveform of the active Mealy FSM operation with the interrupting event Btn

In addition, proposed VHDL models were implemented on FPGA XC3S500E-5fg320. Results of the diagnostic experiment confirmed their correctness and efficiency.

IX. Conclusion

Models of timed control FSMs in logical control systems, which are implemented in hardware, are characterized by a sufficient variety and have a lot of classifications. By the method of output signals generation, FSMs are divided into Moore and Mealy models [11]. By the type and method of input signals procession, FSMs are classified into FSMs that polling input signals and event-driven FSMs; and events, in their turn, are classified as external to the controlled system and internal, while external events also have many classifications [1, 8]. According to states encoding method, FSMs are divided into a number of classes for which hardware costs and speed are differ significantly [7]. According to the method of timing parameters procession for timed FSMs, they are classified into FSMs with timeouts, timing constraints, timing delays and their combinations [3, 5].

By the method of implementing transitions and obtaining output signals, FSMs are classified on regular, timed, and recursive [9]. Such classifications variety of control FSMs, which is implemented in hardware, is due to the wide scope of their application and the variety of tasks to be solved. The method for FSM models classification is decisive during construction of HDL-models patterns for timed control FSMs.

In this work, authors proposed the classification of FSMs into Moore and Mealy models by the method of obtaining output signals, into active and passive models by the method of processing input signals, and the classification of events into initiating and interrupting events by the method of their processing.

This classification made it possible to build VHDL-patterns of timed control FSM models for solving various problems in logical control systems. Implementation in FPGA confirmed that developed patterns belong to the synthesized subset of VHDL and compliance of timing parameters that are specified by specifications.

Further research may be related to the development and analysis of HDL-models using Verilog and the development of HDL patterns for recursive FSMs.

References

[1] Shalyto A.A. Software Automation Design: Algorithmization and Programming of Problems of Logical Control / A.A. Shalyto // Journal of Computer and System Sciences International. – 2000. – Vol. 39, No. 6. – P. 899-916.

[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] Zhigulin M. FSM-Based Test Derivation Strategies for Systems with Time-Outs / M. Zhigulin, N. Yevtushenko, S. Maag, A.R. Cavalli // Proceedings of the 11th International Conference on Quality Software (QSIC 2011), Madrid, 2011. – P. 141-149.

[4] 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'16), October 14-17, Yerevan, Armenia, 2016. – P. 193–196.

[5] Tvardovskii A.S., Yevtushenko N.V., Gromov M.L. Minimizing Finite State Machines with time guards and timeouts // Proc. ISP RAS. – 2017.– vol. 29. – issue 4, – P. 139-154.

[6] Bresolin D. Minimizing Deterministic Timed Finite State Machines / D. Bresolin, A. Tvardovskii, N. Yevtushenko, T. Villa, M. Gromov // In 14th IFAC Workshop on Discrete Event Systems WODES 2018. – IFAC-PapersOnLine, 2018. – Vol. 51. – issue 7. – P. 486-492.

[7] Solov'ev V.V. Structural models of finite-state machines for their implementation on programmable logic devices and systems on chip / A.S.Klimowicz, V.V. Solov'ev // Journal of Computer and Systems Sciences International. – 2015. – V. 54. – № 2. – P. 230-242.

[8] Wagner G. An abstract state machine semantics for discrete event simulation // Proceedings of the 2017 Winter Simulation Conference (WSC), 3-6 Dec. 2017, Las Vegas, USA – 12 p. [Electronic resource] / IEEE Xplore Digital Library – Access mode: www / URL: https://ieeexplore.ieee.org/document/8247830.

[9] Pedroni, V. A. Finite state machines in hardware: theory and design (with VHDL and SystemVerilog) /Volnei A. Pedroni. – Cambridge, MA: MIT Press., 2013. – 338 p.

[10] Shkil A. Design of real-time logic control system on FPGA / M. Miroschnyk, A. Shkil, E. Kulak, D. Rakhlis, I. Filippenko, M. Hoha, M. Malakhov, V. Serhiienko // Proceedings of 2019 IEEE East-West Design & Test Symposium (EWDTS'19), September 13-16, Batumi, Georgia, 2019. – P. 488-491.

[11] Baranov S. Logic and System Design of Digital Systems / S. Baranov. – Tallinn: TUT Press, 2008. – 267 p.

[12] Shkil A.S. Design timed FSM with VHDL Moore pattern / M.A. Miroschnyk, A.S. Shkil, E.N. Kulak, D.Y. Rakhlis, A.M. Miroshnyk, N.V. Malahov // Radio Electronics, Computer Science, Control. – 2020. – №2(53). – P. 137-148.

Modification of VGG Neural Network Architecture for Unimodal and Multimodal Biometrics

Stefanidi Anton
department of infocommunication and radio physics
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
antonstefanidi@mail.ru

Topnikov Artem
department of infocommunication and radio physics
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
topartgroup@gmail.com

Priorov Andrey
department of infocommunication and radio physics
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
andcat@yandex.ru

Kosterin Igor
All-Russian Research Institute for Fire Protection of EMERCOM
Moscow, Russia
kosteriniv@gmail.com

***Abstract*—Identification and authentication systems have become an important part of the modern world. Algorithms for recognizing a person by face, iris, voice, and fingerprint are used in mobile devices, personal data management systems, and banking systems. However, any unimodal biometric system has some inherent disadvantages. In this paper, we consider unimodal and bimodal identification algorithms based on facial and voice biometrics.**

Keywords—multimodal biometric, convolution neural networks, unimodal biometric, speaker identification, face recognition

I. Introduction

Currently, there are a large number of services and applications that use biometric identification methods. Cause this technology allows you to accurately authenticate the user and protect personal data from unauthorized access. As the rule, these are methods of analysis based on a fingerprint, facial image, or voice. However, any unimodal system has some inherent disadvantages. Biometrics based on a fingerprint is an invasive method, which reduces the application areas of this technology. Face recognition systems have a strong dependence on the level of illumination, camera's angle, and quality of the photo recorder, and also they are sensitive to head turns and facial expressions. The speaker identification system depends on the effects of the information channel and microphone, the speaker's physiological characteristics, and the acoustic characteristics of the environment [1-4].

Present study considers the development of a multimodal identification method using facial and voice biometrics. This approach allows you to create a non-invasive system with a high level of protection. Using two biometric parameters reduces the probability data falsification [9, 10]. The first part of this work consists of the analysis of the unimodal method for person identification by using speech signals. In the second part of this research, we consider a bimodal algorithm using a combination of facial and voice biometrics. This work is a continuation of the research [10].

The neural network approach has become one of the main tools in solving problems of object detection, recognition, and segmentation. In particular, methods and algorithms based on neural networks show high results in the people recognition task with using digital images and speech signals [1-4, 11]. Networks are also used in natural language processing, medicine, biochemical research, robotics. Currently, it is really difficult to overestimate the practical importance and need of neural network approaches. This research is also based on the use of convolutional neural networks.

II. Database Description

The VoxCeleb1 database was used in this experiment. It is a modern audiovisual dataset that contains short segments of human speech and faces images. These samples extracted from YouTube video interviews. The total number of celebrities is 1251. The dataset is well balanced by gender. The VoxCeleb1 dataset includes approximately 150000 audio samples. Audio signals were collected at various acoustic conditions and using different types of recording devices. The VoxCeleb1 database also includes a set of face images (Fig. 1). The total number of images is over 1200000. Pictures of faces are taken at different angles and degrees of illumination. Faces have different head tilts, hair and skin color. These celebrities represent people with different ethnicities, accents, professions, and ages [1].

Fig. 1. Examples of face images from the VoxCeleb1 dataset

The VoxCeleb1 is a well-structured dataset. Additionally, the dataset includes face images and audio signals. So VoxCeleb1 is a good variant for creating a multimodal system by using a person's face and voice. We have reduced the number of defined classes from 1251 to 50, because the dataset is quite large and requires computing resources.

III. Data Preprocessing and Description of Network Architecture

Table 1 shows an analyzed part of the audiovisual database VoxCeleb1. The audio signals were in wav format with a sampling rate of 16 kHz and a quantization depth of 16 bits. For standardization, all speech signals were truncated to a duration

of 3 seconds based on random selection of a fragment from the original signal. Each speech sample must be longer than 1 second in the speaker recognition task. We selected a duration of 3 seconds according to [1].

TABLE 1. THE ANALYZED PART OF THE AUDIOVISUAL DATABASE VOXCELEB1

	Train	*Valid*	*Test*	*Total*
Images	43243	2382	2447	48072
Speech signals	5730	271	322	6323

Raw audio data is a change in the amplitude of vibrations over time. It is not an informative form of representation for the speech signal, so in this scientific work we used mel-frequency cepstral coefficients (MFCCs). MFCCs are well suited for representing the speech signal because the mel scale is related to the human's hearing specifics and the frequency resolution of the ear. MFCCs allow you to achieve high results in automatic speech recognition tasks. It is considered good practice to use 40-80 filters. We applied 80 triangular filters (Fig. 2). As a result, each signal was represented by an 80x301 matrix [5].

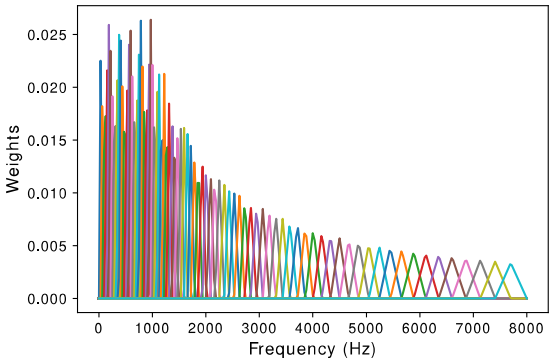


Fig. 2. Bank of 80 triangular mel filters

The method of synthetic data augmentation was used to increase the generalizing ability of trained neural network models. Original speech signals were changed by various transformations: the addition of additive white Gaussian noise; shifting time; changing pitch; time stretch; median-filtering harmonic percussive source separation (HPSS); reverb audio signal. Additionally, Urban Sound Dataset (UrbanSound8K) was used to raise data variability. This dataset contains 8732 labeled sound excerpts less than 4 seconds long of urban sounds from 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. Speech signals were randomly mixed with signals from set the UrbanSound8K [6-8].

A large number of interesting studies based on the analysis of the VoxCeleb1 and VoxCeleb2 datasets are currently published. High results are obtained using large architectures: ResNet18, ResNet34, ResNet50 [1, 2]. They are not suitable for our purpose because we are investigating only a small part of the VoxCeleb1 dataset, which is less than 5% of the total volume. If you apply deep architectures to a small dataset, it is obvious that the network will be overfitting and have weak generalization ability. Therefore, for our research, we designed more compact convolutional neural networks.

Fig. 3 shows the modification of VGG network architecture for speaker identification. We will call this network as CNN-VGGS. This is a network which was used to create a unimodal personality recognition algorithm. The CNN-VGGS gets an mel -frequency cepstral coefficients at the input. The convolutional neural network is very compact since and contains less than 0.3 million weight parameters. For comparison, ResNet18 contains more than 11 million weights, and ResNet50 contains more than 25 million weights.

The CNN-VGGMulti architecture was designed to solve the problem of multimodal identification using two biometric parameters (Fig. 4). This network has two inputs: one for digital facial images with a size of 224x224x3, and the other for MFCCs of speech signals. We note a very important fact about the formation of new features. Each of the threads in the CNN-VGGMulti network has a Global Averaging layer (GA-Pool layer). Vectors of the same 256-dimension are formed at the output of GA-Pool layers. Then these vectors are combined using the concatenation layer Concat into a single common vector of 512-dimensions. This is done so that because the audio and video inputs have the same balanced effect on the final classification result. This network is quite compact and contains 0.95 million weight parameters.

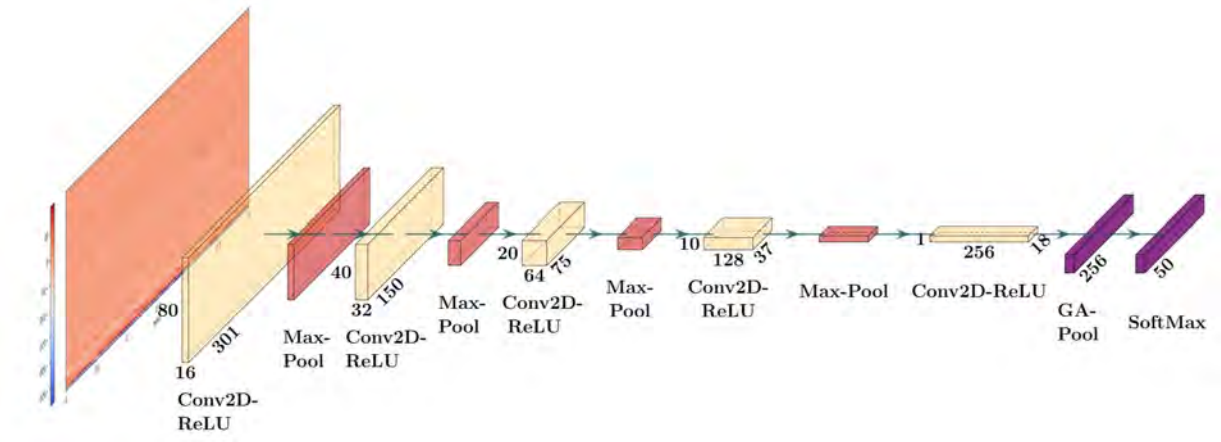


Fig. 3. The architecture of the convolution neural network CNN-VGGS

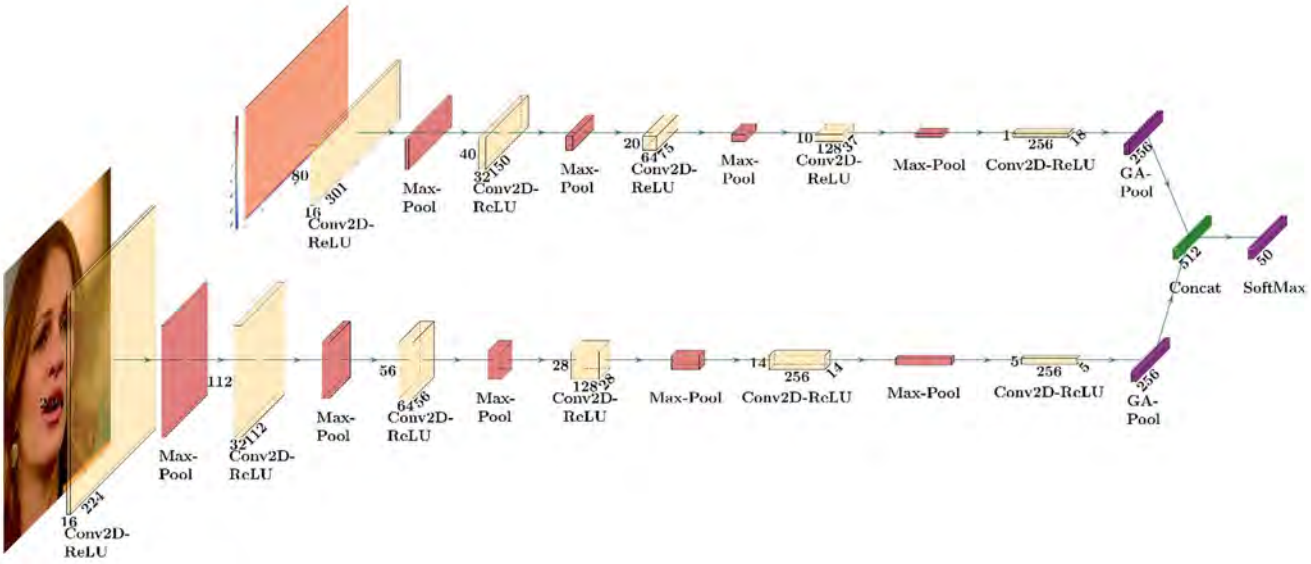


Fig. 4. The architecture of the convolution neural network CNN-VGGMulti

IV. Results

A. Analysis of the Unimodal Algorithm Based on Voice Biometrics

We describe the learning process of the CNN-VGGS neural network and the analysis of test results. Adam (Adaptive Moment Estimation) was used as a method for optimizing weight parameters. During the training the CNN-VGGS, the following hyperparameters were set: optimization algorithm's learning rate was 0.001, the butch size was 32, and the number of epochs was 100. Fig. 5 shows the network learning process using the quality evaluation metric – accuracy (acc).

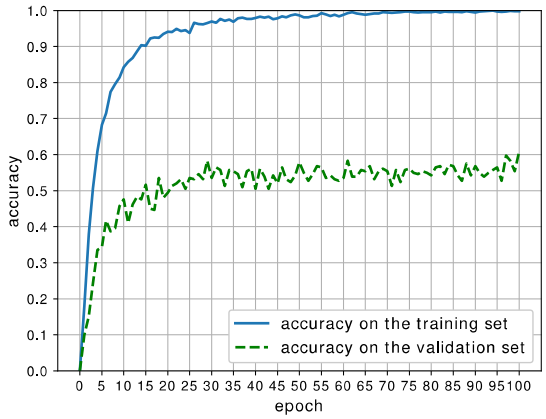


Fig. 5. Accuracy analysis in the CNN-VGGS learning process

The results show that the percent of correct answers on the training dataset is 99.84%, but the score on the verification set is 60.89%, and on the test set is 51.55%. The result indicates that the model is overfitting and it has a low generalizing ability. You can also see that the learning curve is reaching a plateau and increasing the number of epochs will not improve performance.

One of the most widespread methods to prevent with overfitting is the dropout of layers. It is one of the regularization methods. We conducted a number of experiments using this approach but did not get any improvements in the performance. Another common method for increasing the generalization ability model is to increase samples of the training dataset. New samples can be searched on the Internet, or synthetically generated using various types of speech signal transformations. In this research, audio data was changed using different transformations that we described earlier in the current article. Using the method of synthetic augmentation allowed us to slightly improve the score on the test set: 64.95% и 57.76% accordingly. However, the model still has a weak generalizing ability and it is not able to classify the speaker with high precision on new samples.

These results confirm the fact that the speech signal is an extremely difficult representation of information, which very depends on the quality characteristics of recording devices, the level of noise in the sound channel, and the acoustic properties of the environment. Data in the training set has a weak correlation with data from the test set. Also, data augmentation did not significantly improve performance, because the transformations synthesized data that strongly differ in properties from the test set.

We decided to add another biometric feature that well characterizes a person for improving accuracy. As a result, a recognition method was implemented based on a combined analysis of speech signals and face images.

B. Analysis of the Multimodal Algorithm Based on Voice Biometrics

Since the unimodal method of speaker recognition did not show high results on the test set, we decided to develop a multimodal algorithm. For this purpose, the CNN-VGGMulti architecture was implemented. Fig. 6 shows the process of

training a multimodal algorithm based on the CNN-VGGMulti convolutional neural network.

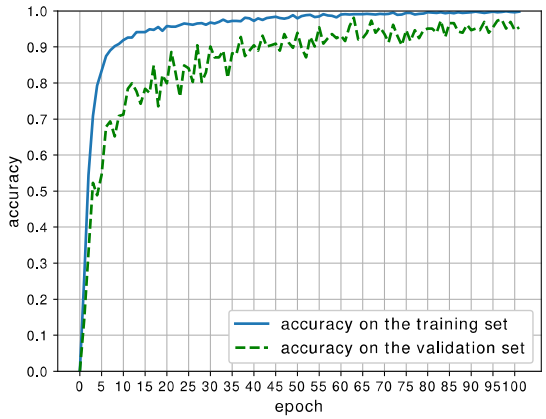


Fig. 6. Accuracy analysis in the CNN-VGGMulti learning process

The neural network includes two branches that have one joint output. The branch for speech data analysis is the CNN-VGGS network. The branch for analyzing face images is also a small convolutional neural network. Each of the branches generates new 256-features based on the corresponding input using the MFCCs or digital face image. Then these features are combined into a common vector with dimension 512-features. The CNN-VGGMulti inputs are tensors of size 80x301x1 and 224x224x3. Adam was used as a method for optimizing weight parameters, by analogy as in the unimodal algorithm. During the training, the following hyperparameters were set: the learning rate of the optimization algorithm was 0.001, the butch size was 8, and the number of epochs was 100.

The multimodal model demonstrates high performance. The score on the training sample is 99.88%, the score on the validation and test datasets is 98.11% and 97.19%, respectively. We can see that the use of a combined approach based on speech and facial biometrics significantly increased the generalizing ability of the classifier. The trained model makes a prediction based on two independent biometric characteristics at the same time. The results of the experiment can be used to design commercial systems for person recognition.

V. Conclusion

The current research was considered the issue of personality recognition using the methods of unimodal and bimodal biometrics. The modern VoxCeleb1 dataset was used as a source of audiovisual information. From the VoxCeleb1 database, 50 unique classes were extracted for the experiment. The research was implemented using only a small part of the Voxceleb1 dataset – around 5%. It allowed the authors to conduct experimental work more dynamically at all steps of the experiment. Analysis of speech signals was based on the extract of Mel-frequency cepstral coefficients.

During the study, two network architectures were designed. The CNN-VGGS network was used for the unimodal person identification based only on speech signals. The research showed that the trained model has a weak generalizing ability and low accuracy of 51.55% on the test set. Dropout in the fully connected layers and synthetic data augmentation were used to prevent overfitting. However, this did not significantly increase the quality of work. The model was trained on the new synthesized data showed an accuracy of 57.76% on the test set.

The recognition algorithm based on the combined analysis of speech signals and digital images was implemented. This algorithm used the architecture of the bidirectional CNN-VGGMulti network. The multimodal model showed high accuracy values. The score on the validation and test dataset is 98.11% and 97.19%, respectively. The results of the research show that the multimodal algorithm can be applied in real recognition systems.

In the next stages of study, we plan to conduct an experiment using the full VoxCeleb1 dataset. We will use the method of combining biometric parameters presented in this paper but with application deeper topologies of ResNet18 and ResNet34 networks to solve the problem of person recognition. Additionally, new methods will be presented for combining biometric parameters.

Acknowledgment

The reported study was funded by Russian Foundation for Basic Research (RFBR), project number 19-37-90158.

References

[1] A. Nagrani, J.S. Chung, A. Zisserman, “VoxCeleb: a large-scale speaker identification dataset,” 2017, Web: https://arxiv.org/abs/1706.08612v2.

[2] J.S. Chung, A. Nagrani, A. Zisserman, “VoxCeleb2: Deep Speaker Recognition,” In Proceedings Interspeech, 2018, pp. 1086-1090.

[3] O.M. Parkhi, A. Vedaldi, A. Zisserman, “Deep face recognition,” In Proceedings British Machine Vision Conference, 2015. pp. 1-12.

[4] Y. Sun, L. Ding, X. Wang, X. Tang, “DeepID3: Face recognition with very deep neural networks,” 2015, Web: https://arxiv.org/abs/1502.00873.

[5] S.K. Kopparapu, M. Laxminarayana, “Choice of Mel filter bank in computing MFCC of a resampled speech. In 10th International Conference on Information Science,” Signal Processing and their Applications, 2010, pp. 121-124.

[6] J. Salamon, J.P. Bello, “Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification,” In IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279-283, March 2017, doi: 10.1109/LSP.2017.2657381.

[7] D.S. Park, W. Chan, Y. Zhang, C.C. Chiu, B. Zoph, E.D. Cubuk and Q.V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” 2019, Web: https://arxiv.org/abs/1904. 08779v3.

[8] J. Salamon, C. Jacoby, J.P. Bello, “A Dataset and Taxonomy for Urban Sound Research. 22nd ACM International Conference on Multimedia,” Orlando USA, Nov. 2014.

[9] V. Khryashchev, A. Topnikov, A. Stefanidi, A. Priorov, “Bimodal person identification using voice data and face images,” In Proceedings SPIE 11041, Eleventh International Conference on Machine Vision, Web: https://doi.org/10.1117/12.2523138.

[10] A. Stefanidi, A. Topnikov, G. Tupitsin, A. Priorov, “Application of convolutional neural networks for multimodal identification task,” 26th Conference of Open Innovations Association FRUCT, 2020, pp 423-428.

[11] V Khryashchev, L Shmaglit, A Shemyakov, “The application of machine learning techniques to real time audience analysis system,” Computer Vision in Control Systems-2, 2015, pp. 49-69.

Development of ICT Models in Area of Safety Education

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

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

Oleksandr Martynyuk
Department of Computer Intelligent Systems and Networks
Odessa National Polytechnic University
Odessa, Ukraine
anmartynyuk@ukr.net

Julia Drozd
Department of Computer Systems
Odessa National Polytechnic University
Odessa, Ukraine
yuliia.drozd@opu.ua

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

***Abstract*—The paper addresses the challenges of developing models that are important for increasing the level of IT security culture in the field of safety-related applications. Models of resource-based approach and peculiarities of their integration into natural world in accordance with vector and levels of resource development are considered. Safety-related systems are analyzed from the position of checkability of digital components and accumulation of hidden faults, which reduce fault tolerance of circuits in emergency mode. The problem of hidden faults was identified as a growth challenge with appropriate ways to address it, including virus epidemics and Covid-19. Case studies of the development of models within the framework of European educational projects aimed at improving the level of culture in IT security are shown.**

Keywords—Culture Issues in IT Security, Resource-Based Approach, Safety-Related System, Digital Component, Checkability, Problem of the Hidden Faults, Growth Challenge, Covid-19

I. Introduction

The artificial world being created by human is becoming increasingly aggressive towards the environment and human himself. Therefore, functional safety and culture issues in IT security are coming to the forefront and becoming the most important. They must be addressed using the highest achievements and, above all, in understanding the problems themselves through the development of our models, that is, our perceptions of evolutionary change.

Problems must be solved resiliently by eliminating their causes, that is, starting with education, raising the level of culture in IT security. It is necessary for ourselves to answer a number of important questions and to use these answers as a basis of education. Why do we want as best but it turns out as always? (V. Chernomyrdin). If the degree of aggressiveness of the artificially created world increases contrary to our aspirations, how great is the objective component of our development, and what is our role in it? What is the vector of our development? How does development happen? What 's good and what 's bad? How are our models reflected in the IT security culture?

To answer these questions, we have to look at the world around us from the outside. Whether it is possible? Turns out that's all we do. Indeed, we study our world in the details, which are its reflection and allow us to draw up about it some integral representation of the observer from the outside.

The rapid development of the computer world creates important prerequisites for the successful improvement of our models. Therefore, the purpose of this paper is to develop ICT models for a better understanding of problems related to functional safety and ways to solve them, which begin in the educational domain with an increase in the level of IT security culture. Section 2 discloses models of a resource-based approach to interpreting evolutionary changes in the natural and computer world by comparing them. Section 3 details these models for critical ICT applications. Section 4 demonstrates the practical results achieved in the educational domain within the models considered.

II. Models of the Resource-Based Approach

Human life is too short to understand the natural world and the designs of its creator. However, we can come substantially closer in this understanding by studying the computer world, as one that repeats the development of the natural world, but in a shorter time frame. Such a view is valid, assuming that the objective component of development dominates and is a single basis for natural and artificial world. This interpretation laid the foundation for a resource-based approach that explores the integration of the computer world into the natural one [1, 2].

The problems addressed can be represented by the productivity with which the entire volume of operations must be performed in the time allocated, the trustworthiness of the results obtained, and the resources invested in achieving the required productivity and trustworthiness. In this sense, resources include everything necessary to solve problems, that is, everything: models, methods and means. As has already been noted, models are our insights into the world around us in its details. The methods ensure development and evaluation of resources. Models and methods form the information component of resources, and means (materials and tools) are their material carriers. Models and methods are written on these carriers in their structure and functioning. For example, the structure of marmot describes the model of future spring which can be fast warm or prolonged cold, and winter hibernation of marmot demonstrates the method of energy saving successfully used in green technologies [3, 4].

978-1-7281-1003-5/19/$31.00 ©2020 IEEE

Productivity and trustworthiness should be interpreted from the perspective of evolutionary understanding of problems as challenges from the natural world, which we ourselves provoke, creating in it perturbations not fully adequate solution to previous integration problems. In this sense, productivity is an assessment of efforts (in time) to solve the problem, and trustworthiness as adequacy to the natural world is the direction of these efforts.

Human mission manifests itself in reading models and methods from the material carriers of the natural world and recording them on the material carriers of the computer world in open code. Between reading and recording, we see the understanding of information resources, their development and verification by creating livelihoods to test models and methods with the practice of using them and with a view to our motivation for their best development. We receive the greatest benefits for the development of the most appropriate models and methods that promote the best integration into the natural world.

Integration takes place by structuring resources under the peculiarities of the natural world according to the method of stick and carrot based on the principles of natural selection and obtaining benefits when guessing the vector of development. From this position, we do not choose the problems that need to be solved and the methods to solve them. With the wrong choice, our place will be occupied by others, and this substitution process will continue until natural selection gives a positive assessment of the choice made. Another result is possible: we know examples of dinosaurs and mammoths that have not received a positive rating. Airships and horse-drawn transport, mechanical calculators and amateur photography and much more, which did not pass another exam for integration, also went into oblivion.

The success of medicine and improvement in the culture of life significantly lowered the threshold of natural selection. Such an artificial buffer between the natural world and humans reduces life forces and causes some concern among scientists. Natural selection is necessary for the survival of the genus, but is realized by the survival (contrary to survival) of its individual representatives, whose interests in the short term are recognized by us as the highest priority. However, the natural world is positioned on the survival side of the genus and for this purpose maintains the threshold of natural selection by seasonal viral epidemics and significantly rarer but also regularly repeated more powerful shakes – pandemics. The usefulness of viral training is due to the fact that viruses are more likely to mutate and cause to mutate other forms of life, i.e. to rise them to a higher level of development for permanent improvement in adaptation to natural changes. Who 's right, who 's to blame? Everyone is right here, and everyone has to do their own thing. The natural world will continue to conduct trainings, and we will try to pass them successfully, including Covid-19. Thus, the natural world is immune from slagging: everything develops in a pair of poison – an antidote. In the artificial world, the same processes are taking place. From accident to accident we become more experienced, and our growing experience – knowledge and skills, i.e. models and methods – elevates us to a higher level of culture in IT security.

The vector of development is determined by the natural world peculiarities, under which resources are structured. The computer world has shown two such features most: parallelism and fuzziness. Its whole history is an example of the permanent increase in the level of parallelism and approximation of solutions. The development of personal computers clearly demonstrates the dominated parallelization of computing in the processing of approximate data. Their hardware-assisted has gone from an optional Intel 287/387 coprocessor to several floating-point pipelines in the Pentium processor and many thousands of such pipelines in the graphics processor used for parallel calculations by CUDA technology. Benefits for following the development vector have been expressed in the remarkable improvement of computers in their productivity and memory. In recent decades, the clock frequency has increased from KHz to MHz, and the amount of memory has increased from MB to TB, that is, the main characteristics have simultaneously improved millions of times [5, 6].

It should be noted that mankind has not built special plans for such development. It came true naturally. It follows that we are led. We are being structured under the peculiarities of the natural world first subconsciously, and then realize the changes perceived by the subconscious, and form our models, within which we develop methods, apply them in practice and receive means. This understanding of the natural world is an evolutionary process that by definition must proceed so slowly that promising resources can adapt.

The resource-based approach identifies three levels in resource development: replication, diversification and self-sufficiency as a development goal. Replication will always dominate in the absence of conflict with the surrounding world, that is, in open resource niches: environmental, technological, market and others. In these circumstances, integration into the natural world is carried out in the simplest way by stamping, cloning with increasing productivity, which should ensure that fertility exceeds mortality. Examples of such survival are demonstrated by vegetative plant propagation methods, bacterial and insect reproduction. However, clones are doomed to extinction when resource niches are being closed. Survival is possible only with the transition to diversification by showing features towards increased adequacy to the natural world, that is, trustworthiness. Self-sufficiency is the goal of development of all resources. To make sure, it is enough to analyse our actions, in which we appear to be a tool for resource development. We strive to create smart resources, give them artificial intelligence to increase self-sufficiency in management, decision-making, communications [7, 8]. We develop the green technologies which are aimed at increasing self-sufficiency in energy consumption [9, 10]. Cloud technology addresses self-sufficiency in performance and memory [11, 12].

Extending the desire for self-sufficiency to all resources, we receive a number of provisions that are important for a common and IT culture of security. They relate to the possibilities of resilient solution of problems, development of risks related to limitations of our models, and overcoming traditional contradictions between parameters of solved tasks.

Problems seek to become independent of the causes. This self-sufficiency creates irreversible processes that impede the resilient, i.e. elastic solution to the problem achieved by eliminating the causes. The solution of self-sufficient problems must be given to the natural world, in the hope of its elasticity and with excluding resistance to the problem. Resistance only increases the problem, and you become its cause. The natural world will solve your problem, but eliminating you as its cause [13].

For example, we often try to shoot down the elevated temperature of our body, fighting not with the disease, but with the protective reaction of the body. The timely elimination of the cause, i.e. when the process is reversible and the problem still depends on the cause, allows the elimination of the disease and the elevated temperature. Resistance to the external manifestation of the disease can significantly reduce the chances of recovery.

Our models are limited by definition. If you represent a model as a circle, its perimeter symbolizes the model 's limitations. Since our models are becoming larger, the perimeter of the circle and these limitations, that is, the risks of misunderstanding of the natural world, are growing.

The parameters of solving any problem also seek self-sufficiency, i.e. independence. This process takes place in our minds, our models as the outlook expands. For example, the relationship between distance, time, and speed, known from a school course, stops working at speeds close to the speed of light. We got used to the idea that for everything it is necessary to pay, for example, for increase in productivity and safety by complication of the decision and increase in energy consumption. In reality, we pay only for misunderstandings of this world, for underdeveloped models. These fees are penalties aimed at encouraging us to develop models.

The contradictions between problem parameters go away as models evolve. Ups and downs on the roads lead to additional costs of gasoline in cars. But the electric vehicle on the descents recharge the battery, compensating for the losses at the ascents, and thus reduces the dependence of energy consumption on the uneven roads. Today, the dynamic power component of digital circuits is determined by the sum of the number of signal transitions from "0" to "1" and from "1" to "0" [14, 15]. But there must be a difference, because transition from "1" to "0" must return energy. In this case, the dynamic component of power consumption will tend to be zero and, for example, circuit testing based on signal transition checking will significantly reduce its dependence on power consumption.

How to develop models? A model has a logical structure folded from concepts, each of which is a terminal element in that structure, i.e. indivisible as an atom. The model develops by splitting these atoms into opposite concepts. This cleavage occurs in our minds in the process of structuring under the realities of the natural world. For example, many diagnostic models of the computer world include the concept of "fault." But a digital circuit fault can be transient or permanent. These two concepts cannot be combined in the real world because the transient fault is a short-term and self-recovering one, but the permanent fault is constantly active. The difference in the faults of two opposing concepts raises us in consciousness to the level of diversification and the model is developed under the slogan of Roman emperors: "Divide et impera."

The natural world does not stamp anything the same, everything develops at the level of diversification. The exception is our models, which generalize many concepts on the basis of conducted analogy and forget about their distinguishing features. This process starts with a model of exact data. We think using exact data, i.e. integer by nature, numbers of elements of sets. Numbering abstracts us from the features of numbered entities and drops us to the replication level. Therefore, the development of our models takes place through diversification of our concepts by structuring under the peculiarities of the natural world. We praise human for his accuracy and consistency. But he must integrate into the world of parallelism and fuzziness.

As a rule, we get initial data for their processing from sensors, that is, we get measurement results, which are approximate data. Already it is necessary to reckon with this fact, at least in view of development of such directions as the Internet of things and everything, the cyber-physical systems, the systems of critical application for which bottom level of their structures are sensors [16-18]. The natural world constantly structures the directions of our development according to its peculiarities.

III. Safety-Related ICT Applications

We have seen an increase in the number and quality of high-risk facilities in energy and transport, space and defense. The increased risk is related to the increased expectation of accidents and is estimated from the position of accident losses and probability of its occurrence [19, 20]. The quantitative growth of high-risk facilities increases density in location and brings them closer to densely populated areas. The qualitative growth of these facilities is evident in their complexity and increase in power. All this increases risks from the point of view of potential losses from possible man-made accidents. No one is going to refuse to develop risky infrastructures. Therefore, risk containment is possible only by reducing the probability of an accident. This mission, aimed at saving the future for all mankind, is entirely based on in-formation technologies, which are being implemented into computer systems, transforming them into safety-related systems [21, 22].

The basis of functional safety is the opposition to failures, i.e. the fault tolerance of the system as the ability to continue to provide services, to perform its functions even in case of failures. But the question arises: "How many failures do you need to consider when designing a fault-tolerant system?" The aircraft must continue normal flight in case of any one failure in the onboard computer. Such a requirement seems justified when there is a negligible probability of two or more failures occurring during a single flight.

However, IT security culture requires special attention to the correctness of assumptions and, in particular, should opposite the substitution of concepts, which occurs when their diversification is insufficient. In this example, the substitution of concepts is the interpretation of the word "occurrence," since it is not the occurrence but the manifestation of faults that is important for fault tolerance. The three-channel majority system is fault tolerant in case of faults in each channel, if only one of them appears as an error. The distinction between the meaning of the words "occurrence" and "manifestation" leads to the search for sources of multiple faults that may arise before the flight, and appear already in its process. The source of such faults of the system is its insufficient checkability, creating conditions for hidden processes and accumulation of hidden faults. In the test mode, digital circuits are characterized by testability, which is a structural checkability dependent only on the structure of the circuit [23, 24]. On-line testing makes the circuit checkability dependent also on input data [25, 26].

Hidden faults are completely harmless in conventional computers operating in the same operating mode, and become a problem for today's safety-related systems. All such systems

rise to the level of diversification, which manifests itself in the division of the operating mode into normal and emergency ones. For modern systems, this process is accompanied by data diversification, which become different in these modes at the inputs of digital components due to their traditional design at the replication level, i.e. based on matrix structures for processing data in parallel codes. Different input data lead to diversification of the system in its checkability, which becomes different in normal and emergency mode and thus creates a problem of hidden faults. They can be accumulated in normal mode in the absence of their input data and appear in emergency mode in quantity exceeding the capabilities of fault-tolerant solutions [27-29].

The resource approach identifies the problem of hidden faults as a challenge of growth: the system rises to the level of diversification, and its components continue to be stamped at the level of replication. This interpretation of the problem is a significant contribution to the IT security culture, as it allows it to consciously address the growth challenge by purposefully developing components to the system level [13].

At present, this problem is solved using accident imitation modes, which have repeatedly led to emergency consequences by their unauthorized activation or as a result of deliberate blocking of emergency protections at planned activation [30, 31]. At the same time, the development of components to the system level can be carried out during their design by equalizing the checkability of circuits in normal and emergency mode, for example, with the transition to bit-by-bit pipeline processing [32].

The development of viral epidemics and pandemics, including Covid-19, is also characterized by insufficient checkability when viruses exploit the possibility of hidden propagation at the replication level. Therefore, the main method of countering epidemics is to limit and close resource niches by introducing quarantine. Isolation eliminates the virus in two ways: recovery or death. Medical scientists also fear a third way when the virus mutates during the epidemic and could invent new ways of spreading at the diversification level [33].

It should also be noted that viral epidemics, much less pandemics such as Covid-19, are also identified and addressed as a growth challenge. They start with a mutation that raises the virus to a level of diversification. In this case, protections, including tests and drugs procured by replication for a previous version of the virus, become obsolete and need to be developed to the level of mutated virus.

IV. Case Studies

Models of resource-based approach and the culture of IT security began to be developed within the projects including TEMPUS SAFEGUARD «National Safeware Engineering Network of Centres of Innovative Academia-Industry Handshaking» (158886-TEMPUS-1-2009-1-UK-TEMPUS-JPCR) and TEMPUS GREENCO «Green Computing & Communications» (530270-TEMPUS-1-2012-1-UK-TEMPUS-JPCR), ЕРАЗМУС+ ALIOT «Internet of Things» (573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP) which are carried out since 2009 till present with assistance of the European educational programs of TEMPUS and ERASMUS. The projects focused on the development of master's and postgraduate courses on functional safety of critical applications, green technologies and the Internet of Things. Work with the projects coordinated by the University of Newcastle from the European Union and National Aerospace University "Kharkiv Aviation Institute" (Department of Computer Systems, Networks and Cybersecurity) in Ukraine.

Developed master's courses have been and continue to be read at project consortium universities including Odessa National Polytechnic University. Laboratory practicums delivered as part of the courses introduce students to the practical side of resource approach models and IT security culture. Laboratory practicums have been developed with student participation based on software and hardware models of arithmetic units. Software models are developed in a Delphi environment on a freely distributed demo-version. Hardware models are designed using FPGA technology in Quartus CAD.

The most common subject of research is the iterative array multiplier of mantissas. This object is selected for a number of reasons. First, the multiplier implements a key operation for approximate data that is processed typically in floating-point formats where multiplication is used in the number representation itself [34, 35]. For this reason, multiplication in one form or another is present in all operations with mantissas, and the results of these operations inherit the properties of the product. Second, the iterative array multiplier is widespread and is a bright representative of objects traditionally designed at the replication level. So, we can use an iterative array multiplier to investigate the question: "Designing at the replication level is good or bad?"

Students study an A-bit iterative array multiplier that contains A^2 operating elements and performs the operation in one clock cycle. In the fastest scheme, the duration of the clock cycle is determined by serial connection of the $2(A-1)$ operational elements. Therefore, each operating element is used in the clock cycle only for its small part $B = 0.5 / (A-1)$. For example, for $A = 32$, each of a thousand of operational elements is used by $B = 1.6\%$. What happens to the iterative array multiplier with transition to $A = 64$? The number of operational elements is increased by four times, and their usability is reduced twice to $B = 0.8\%$.

The study of the iterative array multiplier continues in a laboratory practicum that counts the number of functional and parasitic transitions in a single operation. Parasitic transitions are caused by the competition of signals propagating along paths of different lengths [36]. The number of transitions determines the dynamic component of power consumption. Parasitic transition is known to increase the total power consumption of an 8-bit adder by 30% [37]. The iterative array multiplier program model shows that functional transitions occur at an average of 30% of circuit points, and the number of parasitic transitions exceed the number of functional transitions by $C = 3.5$ times for $A = 8$. This excess increases with the bit size of the iterative array multiplier and is $C = 4.5$ and $C = 5.5$ for $A = 10$ and $A = 12$, respectively. Thus, the dynamic component of power consumption is mainly determined by parasitic signal transitions, as well as the static component is defined by large matrix sizes.

The next laboratory practicum examines the dual-mode checkability of the iterative array multiplier in a safety-related application by determining the number D of potentially dangerous points where a hidden fault may appear. Checkability E is defined as an addition to one of percentage of these points.

Figure 1 shows the main panel of the program model, which reveals the structure of the iterative array multiplier and highlights the operating elements containing potentially dangerous points in yellow, as well as their number.

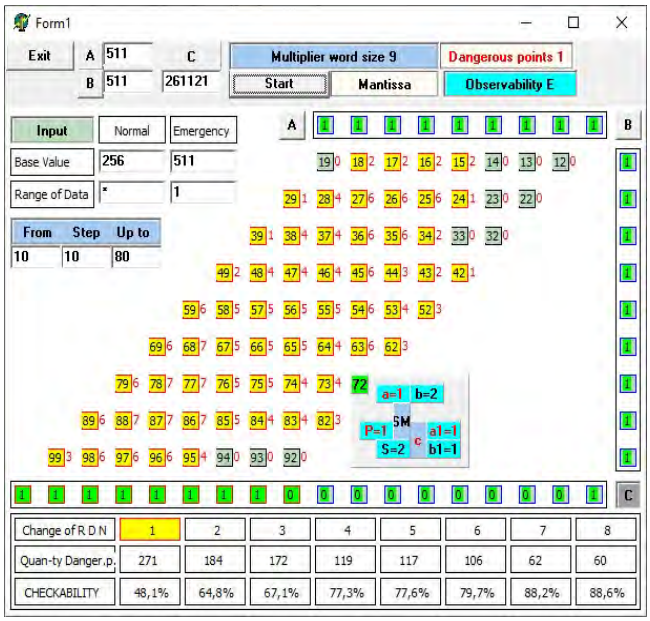


Fig. 1. Iterative array multiplier checkability study

The program calculates parameters D and E for eight experiments with a different number of normal mode input words. In the first experiment, multiplicands take ten values ($G = 10$) and form $G^2 = 100$ normal mode input words. In the following experiments, the number of values received by the multiplicands is increased with step $H = 10$ and the number of input words is increased as follows: 400, 900, 1600, 2500, 3600, 4900 and 6400. Experiments show how the number of potentially dangerous points increases, and the checkability of the circuit reduces with a decrease in the number of input words in normal mode. The large number of potentially dangerous points and the low checkability of the iterative array multiplier characterize it as a carrier of the matrix structure.

The laboratory practicums described introduce students to the problems of matrix structures that reflect replication as a lower level of resource development. Problems are shown in terms of low operational element utilization, irrational power consumption, and limited checkability of the matrix scheme in safety-related applications.

The next round of laboratory practicums explores on-line testing methods. Traditional methods such as residue checking the complete arithmetic operations have been developed in the theory of construction of totally self-checking circuits, which developed within the model of exact data. The main requirement for these methods is to detect faults from a given set using the first erroneous result [38].

The approximate result diversifies its bits into most and least significant bits [39]. In these bits, faults cause errors that are respectively essential and inessential to the trustworthiness of the result. Therefore, the trustworthiness of on-line testing methods is determined by the probabilities of detecting essential errors and skipping inessential ones. Experiments show low trustworthiness of traditional methods due to more frequent occurrence of inessential errors than essential ones.

Laboratory practicums use Fault Injection Technology [40] to evaluate the trustworthiness of on-line testing methods under conditions of stuck-at and shorts faults [41]. Traditional methods are compared with on-line testing methods aimed at processing approximate data [42, 43]. Mantissa processing is based on the methods of truncated execution of operations [44-46], which almost halve the expense of equipment and time without loss in accuracy, by following the development vector.

Increasing the level of circuit parallelism traditionally aims to increase productivity in performing calculations. The following laboratory practicum shows another, more important goal: to follow the development vector on the example of an arithmetic shifter of mantissa as a private case of an iterative array multiplier. The program model allows to investigate the shift at several levels of circuit parallelism and with its increase shows a multiple effect of improvement of indicators: not only growth of productivity / complexity ratio, but also at the same time gain in trustworthiness of results and multiple reduction of dynamic component of energy consumption.

Splitting of the concept "fault" on two contrasts: transient and permanent one is shown in a laboratory practicum which investigates restoration of the erroneous results in cases of transient and permanent faults for digital delays. Taking into account the single manifestations of transient fault and the permanent nature of failure, the correct result is obtained at a lower cost than the majority system, which does not distinguish between these cases.

V. Conclusions

Education should reflect advanced development models, and this is particularly true of IT security culture and safety-related applications in general. Their and our functional safety, which has long become unified, is primarily determined by the level of culture in safety and, accordingly, by models and methods implemented in safety-related systems. The increased risk of critical applications, due in large part to the lack of checkability of systems and their components, presents a problem of hidden faults. This problem creates distrust of the fault tolerant solutions underlying functional safety and pushes to use dangerous imitation modes, that recreate emergency conditions to detect hidden faults. A study of resource levels shows that the problem of hidden faults is a challenge of growth, when the system has already reached the level of diversification and its components continue to be stamped at the level of replication. This understanding of the problem allows it to be solved as a challenge of growth by raising components to the level of the system.

The vector of development demonstrated by the computer world manifests itself in its structuring under the parallelism and fuzziness of the natural world. Models and methods develop in our minds by diversifying the concepts used. This process takes place naturally following our subconscious, which is shaped by the realities of the world around us. However, knowledge of the development vector and its support in education allows to actively accompany natural processes, thus increasing the level of safety culture and consciously promoting successful integration into the natural world.

The natural world permanently teaches us lessons. Their study should be supported by education, which is obliged to repeat these lessons in lecture courses and laboratory practicums, increasing the level of culture in IT security.

REFERENCES

[1] J. Drozd, A. Drozd, S. Antoshchuk, "Green IT engineering in the view of resource-based approach," in book: Green IT Engineering: Concepts, Models, Complex Systems Architectures, SSDC, vol. 74. Springer, Berlin, 2017, pp. 43-65. DOI: 10.1007/978-3-319-44162-7_3

[2] J. Drozd, A. Drozd, M. Al-dhabi, "A resource approach to on-line testing of computing circuits," IEEE East-West Design & Test Symposium, Batumi, Georgia, 2015, pp. 276-281. DOI: 10.1109/EWDTS.2015.7493122

[3] L. Saker, S. Elayoubi, T. Chahed, "Minimizing Energy Consumption via Sleep Mode in Green Base Station," IEEE Wireless Communication and Networking Conference, Sydney, Australia, 2010.

[4] C. Qiu, C. Zhao, F. Xu, T. Yang, "Sleeping mode of multi-controller in green software-defined networking," EURASIP Journal on Wireless Communications and Networking 2016, Article 282.

[5] S. Chernov, S. Titov, L. Chernova et. al., "Algorithm for the simplification of solution to discrete optimization problems," Eastern-European Journal of Enterprise Technologies, 2018, 3 (4), pp. 1-12.

[6] K. Asanovic, R. Bodik, J. Demmel et. al., "A view of the parallel computing landscape". Commun. ACM, 2009, 52 (10): pp. 56-67. DOI:10.1145/1562764.1562783

[7] Alice Pavaloiu, "The Impact of Artificial Intelligence on Global Trends," Journal of Multidisciplinary Developments. 2016, vol. 1, issue 1, pp. 21-37.

[8] Meenakshi Nadimpalli, "Artificial Intelligence Risks and Benefits," International Journal of Innovative Research in Science, Engineering and Technology, June 2017, vol. 6, issue 6.

[9] Harnessing green IT: principles and practices / San Murugesan, G. R. Gangadharan (edits), John Wiley and Sons Ltd, 2012.

[10] 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.

[11] B. Sisisky, Cloud Computing Bible, John Wiley & Sons, January, 2011.

[12] V. Hahanov, A. Zhalilo, W. Gharibi, E. Litvinova, "Cloud-driven traffic control: Formal modeling and technical realization," 4th Mediterranean Conference on Embedded Computing, MECO, 2015. pp. 21-24.

[13] O. Drozd, V. Kharchenko, A. Rucinski et. al., "Development of Models in Resilient Computing," IEEE International Conference DESSERT, Leeds, UK, 2019, pp. 2-7. DOI: 10.1109/DESSERT.2019.8770035

[14] S. Mittal, "A survey of techniques for improving energy efficiency in embedded computing systems", IJCAET, 2014, 6(4), pp. 440-459.

[15] S. Khatamifard, L. Wang, A. Das et. al., "POWERT channels: a novel class of covert communication exploiting power management vulnerabilities," International Symposium on HPCA, Washington, USA, 2019, pp. 291-303.

[16] D. Maevsky, A. Bojko, E. Maevskaya et. al., "Internet of things: Hierarhy of smart systems," 9th IEEE International Conference IDAACS, Bucharest, Romania, 2017, pp. 821-827.

[17] V. Hahanov, A. Hahanova, S. Chumachenko, S. Galagan, "Diagnosis and repair method of SoC memory," WSEAS Transactions on Circuits and Systems, 2008, vol. 7, nu. 7, pp 698-707.

[18] D. Smith, K. Simpson, The Safety Critical Systems Handbook, 5th Edition, Butterworth-Heinemann, 2019.

[19] S. Choe, F. Leite, "Assessing safety risk among different construction trades: Quantitative approach," Journal of Construction Engineering and Management, 2017, vol. 143, issue 5, 04016133.

[20] O. Ivanchenko, V. Kharchenko et. al., "Risk assessment of critical energy infrastructure considering physical and cyber assets: methodology and models," IEEE 4th International Symposium IDAACS-SWS, 2018.

[21] IEC 61508-1:2010. Functional Safety of Electrical / Electronic / Programmable Electronic Safety Related Systems – Part 1: General requirements. Geneva: IEC, 2010.

[22] Core Knowledge on Instrumentation and Control Systems in Nuclear Power Plants, Technical Reports, IAEA Nuclear Energy Series No. NP-T-3.12, International Atomic Energy Agency Vienna, 2011.

[23] V. Romankevich, "Self-testing of multiprocessor systems with regular diagnostic connections," Automation and Remote Control, 2017, vol. 78, issue 2, pp. 289-299.

[24] K. Zashcholkin, O. Drozd, "The Detection Method of Probable Areas of Hardware Trojans Location in FPGA-based Components of Safety-Critical Systems," IEEE International Conference DESSERT, Kyiv, 2018, pp. 212-217. DOI: 10.1109/DESSERT.2018.8409130

[25] A. Drozd, J. Drozd, S. Antoshchuk et. al., "Objects and Methods of On-Line Testing: Main Requirements and Perspectives of Development," IEEE East-West Design & Test Symposium, Yerevan, Armenia, 2016, pp. 72-76. DOI: 10.1109/EWDTS.2016.7807750

[26] 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.

[27] O. Drozd, M. Kuznietsov, O. Martynyuk, M. Drozd, "A method of the hidden faults elimination in FPGA projects for the critical applications," 9th IEEE International Conference DESSERT, Kyiv, Ukraine, 2018, pp. 231-234. DOI: 10.1109/DESSERT.2018.8409131

[28] A. Drozd, S. Antoshchuk, J. Drozd et. al., "Checkable FPGA Design: Energy Consumption, Throughput and Trustworthiness," in book: Green IT Engineering: Social, Business and Industrial Applications, SSDC, vol. 171, Springer, Berlin, 2019, pp. 73-94. DOI: 10.1007/978-3-030-00253-4_4

[29] 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.

[30] D. Gillis, "The apocalypses that might have been," DAMN Interesting, No 298, 2007. https://www.damninteresting.com/the-apocalypses-that-might-have-been/

[31] E. Blakemore, "The Chernobyl disaster: What happened, and the long-term impacts," 2019 [Online]. Available: https://www .nationalgeographic.com/culture/topics/reference/chernobyl-disaster/

[32] O. Drozd, V. Antoniuk, V. Nikul, M. Drozd, "Hidden faults in FPGA-built digital components of safety-related systems," 14th International Conference "TCSET'2018, Lviv-Slavsko, Ukraine, 2018, pp. 805-809. DOI: 10.1109/TCSET.2018.8336320

[33] N. Grubaugh, M. Petrone, E. Holmes, "We shouldn't worry when a virus mutates during disease outbreaks," Nature Microbiology, 2020, 5(4).

[34] 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.

[35] Synopsys. DWFC Flexible Floating Point Overview, no. August, pp. 1-6, 2016.

[36] V. Kumar, M. Sharma, K. Lal Kishore, A. Rajakumari, "Selective glitch reduction technique for minimizing peak dynamic IR drop," Microelectronics and Solid State Electronics 2013, 2(2A), pp. 27-32

[37] A. Chandracasan, R. Sheng, S. Brodersen, "Low-Power CMOS Digital Design," IEEE Journal of solid-state circuits, 1992, vol. 27, no. 4, pp. 473-484.

[38] C. Metra, L. Schiano, M. Favalli, B. Ricco, "Self-checking scheme for the on-line testing of power supply noise," Design, Automation and Test in Europe Conference, Paris, France, 2002, pp. 832-836.

[39] S. Akinola, 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.

[40] M. Le, A. Gallagher, Y. Tamir, "Challenges and Opportunities with Fault Injection in Virtualized Systems," International Workshop on Virtualization Performance: Analysis, Characterization, and Tools, Austin, Texas, 2008.

[41] W. Pleskacz, M. Jenihhin, J. Raik et. al., "Hierarchical Analysis of Short Defects between Metal Lines in CMOS IC," 11th Euromicro Conference DSD, Parma, Italy, 2008, pp. 729-734.

[42] A. Drozd, S. Antoshchuk, "New on-line testing methods for approximate data processing in the computing circuits," 6th IEEE International Conference IDAACS, Prague, Czech Republic, 2011, pp. 291-294. DOI:10.1109/IDAACS.2011.6072759

[43] A. Drozd, M. Lobachev, "Efficient on-line testing method for floating-point adder," Design, Automation and Test in Europe. Conference and Exhibition 2001 (DATE 2001), Munich, Germany, 2001, pp. 307-311. DOI: 10.1109/DATE.2001.915042

[44] H. Park, "Truncated Multiplications and Divisions for the Negative Two's Complement Number System," Ph.D. Dissertation. The University of Texas at Austin, Austin, USA, 2007.

[45] V. Garofalo, "Truncated Binary Multipliers with Minimum Mean Square Error: Analytical Characterization, Circuit Implementation and Applications," Ph.D. Dissertation. University of Studies of Naples "Federico II", Naples, Italy, 2008.

[46] A. Drozd, M. Lobachev, W. Hassonah, "Hardware check of arithmetic devices with abridged execution of operations," European Design & Test Conference, Paris, France, 1996, P. 611. DOI: 10.1109/EDTC.1996.494375

Big Data Critical Computing Based on the Similarity-Difference Metric

Abdullayev Vugar Hacimahmud
Computer Engineering Department
Azerbaijan State Oil and Industry University
Baku, Azerbaijan
abdulvugar@mail.ru

Lyudmila Shapa
Design Automation Department
Kharkov National University of Radio Electronics
Kharkov, Ukraine
d_ad@nure.ua

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

Alexander Mishchenko
Design Automation Department
Kharkov National University of Radioelectronics
Kharkov, Ukraine,USA
santific@gmail.com

Olga Shevchenko
Design Automation Department
Kharkov National University of Radio Electronics
Kharkov, Ukraine
olga.shevchenko@nure.ua

Svetlana Chumachenko
Design Automation Department
Kharkov National University of Radio Electronics
Kharkov, Ukraine
svetachumachenko@icloud.com

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

***Abstract*— Models, methods and algorithms for cyber-social computing are proposed that use the similarity-difference metric of unitary coded information for processing big data in order to generate adequate actuator signals for controlling cyber-social critical systems. A set-theoretic method for data retrieval has being developed based on the similarity–difference of the frequency parameters of primitive elements, which makes it possible to determine the similarity of objects, the strategy of transforming one object into another, as well as to identify the level of common interests, conflict. The definitions of the fundamental concepts in the field of computing are given on the basis of metric relations between interacting processes and phenomena. A software application is proposed for calculating the similarity-differences of objects based on the formation of frequency vectors of two sets of primitive data. A high level of correlation between the results of the application and the well-known system for determining plagiarism is shown.**

Keywords— computing, cyber social computing, decision-making, unitary data codes, similarity–difference, big data analysis, plagiarism

I. MOTIVATION AND DEFINING RESEARCH OBJECTIVES

A critical system is a set of relations (integrity and unity) interconnected in cyber-physical space and time between the components to achieve the set goal, the failures of which lead to significant economic, political, social, environmental and humanitarian (material-energy and space-time) losses. Examples of critical systems [1] are technological and technical objects in the following fields: energy, transport, industry, weapons, cyber-social sphere, banking, Internet, statehood, and jurisprudence. Unambiguously, scientists and specialists have made the conclusion that about 80 percent of all failures in critical systems are associated with a person's inability to control any systems or objects, including himself [1, 2]. Human is always only a bad performer. Consequently, it is necessary to exclude him from the cycle of monitoring and control, at least of critical processes and phenomena, by transferring the decision-making authority to deterministic and practically error-free computing: network, cloud, terminal [3, 4]. The winner is the one who timely transforms the physical and social space into digitized processes and phenomena for accurate monitoring and control, (preferably human-free). The most unreliable critical system is the authoritarian statehood of an incompetent immoral leader, which is capable of sacrificing the lives and well-being of millions of citizens, the country's economy and ecology. Therefore, personnel management in critical systems of any nature remains the most important problem of humanity, solving of which is associated with the preservation of a planet suitable for human life. Computer engineering is a branch of knowledge that involves the theory and practice of design, testing, production and operation of secure software and hardware scalable computing applications for reliable metric management of virtual, physical and social processes and phenomena by using intelligent cloud and telecommunication services based on digital monitoring of cyber-physical space using personal gadgets and built-in smart sensors. Here, computing, as a global methodology that can be used in computer engineering too, is a strategy for achieving and visualizing a set goal – creating products and/or services with given resources, which is systematically represented by the processes of monitoring and actuation of metric relations in a closed infrastructure of management and execution. Computing can be systematically represented (Fig. 1) by the process of monitoring (5) and actuation (6) metric relations (2) in the infrastructure of management (3) and execution (4) to achieve and visualize (8) the set goal – products and/or services (1) for given resources (7).

The metric and structural definition of computing through eight interrelated components provides a theoretical fundamental basis for the formal and actual creation of a digital control system for any process in a given area of human or natural activity. The types of computing according to the

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

introduced metric cover all fields of human activity: cosmological, biological, floristic, physical, virtual, quantum, social, state, medical, transport, infrastructural, scientific, educational, industrial, sports, recreation, travel, entertainment. Naturally, computing is primarily focused on human-free monitoring and control of critical objects, processes and phenomena.

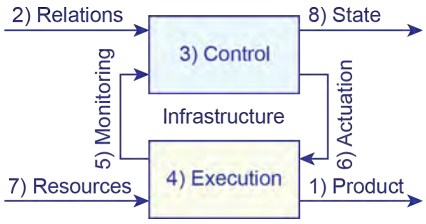


Fig. 1. Critical HR management system computing

Citizens' mobility gives rise to interesting alternative proposals from existing states, which are increasingly competing for the planet's human resources according to the following metric: the smartest and cheapest. Here, each person also acquires the right to make an alternative choice of the state (Fig. 2), which forms the quality of life for an employee according to the metric of relations: salary level, language, culture, traditions, history, food, infrastructure, transport, climate, political stability, social benefits, healthcare, tax legislation. It is obvious that financial flows from citizens to alternative states are directly proportional to the above-mentioned metric of relations on the part of the authorities to a person. Today, at least two or more states will participate in the competition for an employee. The one that offers the best conditions for life and creativity wins. The other gradually self-destructs.

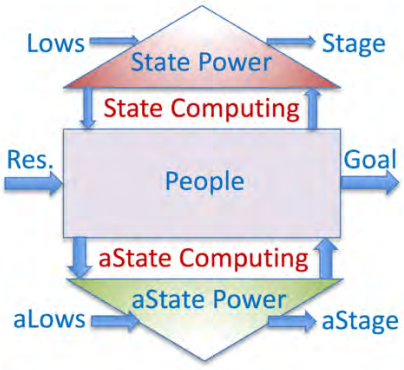


Fig. 2. An alternative to the weak statehood

Competitive, strictly metric, interactive computing on the market for talented workers and statehoods is emerging, where every citizen of the planet chooses the state that is most favorable for creativity and recreation, and the social system chooses the best, creative and healthy performers. The quality of a product (service, process or phenomenon) is a set of properties that determine its suitability to meet certain needs in accordance with its purpose. The quality metric of a critical system is determined by the following parameters: reliability, durability, maintainability, preservation, testability, controllability, observability, diagnosability, serviceability, safety and survivability. As for critical situations and failures, at present there is comprehensive information in the cyber-physical space about any negative process or phenomenon that can be prevented by means of intelligent cloud and edge computing of monitoring and control, which is the essence of critical system computing (Fig. 3). Here, two computers (cloud and terminal) serve the critical system using sensors and actuators. Naturally, cloud computing is invariant with respect to the time and location of a critical object, for example, a car (Synopsys, GMC, Tesla). Quality and reliability are ensured here by the standards: JTAG IEEE 11.49, SECT IEEE 1500, IJTAG IEEE 1687, ISO 9001. Boundary scanning technologies of the aforementioned standards create additional lines and spare components that allow to achieve high levels of quality and reliability through online testing and repairing critical systems using built-in BIST tools and cloud test services.

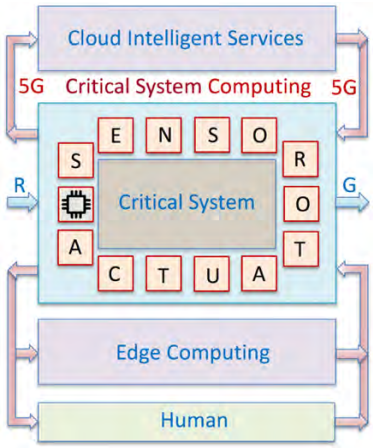


Fig. 3. Critical system computing

Naturally, only a competent operator has the ability to interfere with the operation of a critical system through a terminal computer. To do this, it is necessary to perform an exhaustive monitoring of the competencies of each person by using retrieval and special applications, in order to subsequently adopt of actuator decisions on the appointment of an employee to a functional position, which is the essence of personnel computing or HR–Management. To measure the competencies of employees, represented by vectors of variables, the Levenstein metric is used, which makes it possible to determine the similarity-difference between applicants and a reference pattern, as well as quasi-optimal routes for transforming one metric-model into another.

II. The Purpose and Objectives of the Research

The goal of the research is to reduce the economic, technological and social losses associated with minimizing failures in critical systems by increasing the competence of employees and the consistent exclusion of human from decision-making processes based on his replacement by deterministic computing mechanisms using digital intelligent control based on metric monitoring of cyber-social processes and phenomena.

The objective function L is minimization of direct D and indirect losses S associated with n-failures and repair of R critical systems, including the cost of developing and maintaining computing structures of metric online decision-making on digital control of critical processes and phenomena

based on exhaustive, accurate monitoring M, using smart infrastructure I and qualified employees E meeting the reference competencies in education, experience and skills:

$$L=min\,\Sigma_{i=1}^{n}(D_i + k_i \times S_i + R_i) \leftarrow (A + M + I + E) \leq G_{min}.$$

Objectives to be solved in order to achieve the goal: 1) Develop a structural model of computing for interactive online interaction between human, a critical system and precise digital monitoring-control mechanisms. 2) Develop a theory and data structures for the frequency-multiple method for determining the similarity of two objects. 3) Synthesize a similarity-difference algorithm for text fragments. 4) Perform testing and verification of the method using examples.

III. Frequency-multiple method for determining the similarity of two objects

The technological core for solving practical problems of personnel management in critical systems is cyber-physical computing, formalized in the structure of Machine Learning and SCADA – Supervisory Control And Data Acquisition. Such computing requires big data analytics that use primitive set-theoretic operations, procedures, and parallel algorithms to improve productivity in finding quasi-optimal solutions. Therefore, further it is proposed to implement the algorithm and procedures in the software code to retrieve data according to a given pattern by comparison, which makes it possible to take adequate management actions in critical systems [4]. In the world of retrieval-computing, there is nothing but a similarity-difference metric [5-8]. Therefore, it is important to have an effective specialized processor as the simplest core for the parallel and high-performance solving problems of synthesis and analysis of new processes and phenomena. Structurally, the similarity-difference metric of two processes, phenomena, objects, components uses two formulas, operating in binary algebra of logic with two parallel operations and, xor to obtain the resulting vectors [9]:

$$S(a,b) = a_{i\ \underset{i=1,n}{\wedge}}\ b_i;$$
$$D(a,b) = a_{i\ \underset{i=1,n}{\oplus}}\ b_i.$$

But such formulas are not very effective for identifying relationships between processes (phenomena), when it is necessary and very important to determine common data structures in order to understand how individual components (vector coordinates) transform into each other during synthesis and analysis. Moreover, here the entire synthesis process is computationally dependent on technologically advanced data structures. The normalized similarity-difference metric uses two formulas, which also operate in the algebra of logic by two parallel operations but supplemented by the arithmetic of calculating unit coordinates obtained as a result of performing logical operations. In addition, a common denominator appears in the form of a disjunction of the same coordinates of vectors, which is an integrator of disparate data structures of the considered processes into a common vector of precisely and only essential coordinates, relative to which the similarity and difference are normalized:

$$S(a,b) = \frac{\Sigma_{i=1}^{n}\left(a_{i\ \underset{i=1,n}{\wedge}}\ b_i\right)}{\Sigma_{i=1}^{n}\left(a_{i\ \underset{i=1,n}{\vee}}\ b_i\right)};$$
$$D(a,b) = \frac{\Sigma_{i=1}^{n}\left(a_{i\ \underset{i=1,n}{\oplus}}\ b_i\right)}{\Sigma_{i=1}^{n}\left(a_{i\ \underset{i=1,n}{\vee}}\ b_i\right)}.$$

For instance, two vectors a= 00111100 and b= 10101010, having insignificant zero coordinates of the same name are automatically excluded from the normalized estimation through taking into account and counting only unit values in the resulting vectors:

$$S(a,b) = \frac{\Sigma_{i=1}^{n}(00111100\wedge 10101010=001010000)=2}{\Sigma_{i=1}^{n}(00111100\vee 10101010=10111110)=6} = 0{,}33;$$
$$D(a,b) = \frac{\Sigma_{i=1}^{n}(00111100\oplus 10101010=10010110)=4}{\Sigma_{i=1}^{n}(00111100\vee 10101010=10111110)=6} = 0{,}66;$$

Naturally, there is no need to calculate both estimates using these formulas. It is enough to define one of them, and the second can be obtained by the complement formula:

$$D(a,b) = 1 - S(a,b);\ S(a,b) = 1 - D(a,b).$$

Here, the difference between the formed estimate and the Hamming distance is the exclusion from the metric and data structures of the condition of existence of two zeros on coordinates with the same address-indexes, which significantly increases the adequacy of the measurement of two processes. As for multivalued algebra (set theory), where symbols, letters, numbers, words, texts, objects, processes are used instead of the alphabet {0,1}, the similarity-difference is usually considered within the framework of the Levenstein metric or distance. It involves three elementary operations: character replacement, insertion and deletion, which transform one word (process, phenomenon) into another. Another solution is proposed for determining the similarity between words, which is characterized by the synthesis of a unified data structure that aligns pairs of words of any length to one dimension by performing a single operation – inserting a blank (empty) character. As a consequence, the computational complexity of the algorithm for synthesizing a unified structure of a single dimension is reduced to finding positions to insert a finite number n = 0,1,2,3, ... blank symbols in order to align the length of two words (objects, processes). As an example, the following transformation of one word to another by inserting blank characters is shown:

C O N D U C T I O N
B O N D I A N A

Executing the algorithm for inserting blank characters in order to obtain the minimum difference and maximum similarity when transforming one word into another gives the following result:

C O N D U C T I O N –
B O N D – – – I A N A

The number of blank characters to align two words is four. After that, a trivial calculation of the Levenstein distance is carried out, which is equal to the number of coordinates having different symbols in the metric of word transformation, which means $D(a,b) = 6$, $S(a,b) = 5$. Thus, any pair of processes or phenomena can be reduced to a structural metric of the same

length in order to subsequently calculate normalized similarity-difference scores by the arithmetic addition of the fulfillment of logical conditions in the numerator and denominator:

$$S(a,b) = \frac{\Sigma_{i=1}^{n}(a_i = b_i)}{\Sigma_{i=1}^{n}(a_i \cup b_i \neq \emptyset)};$$
$$D(a,b) = \frac{\Sigma_{i=1}^{n}(a_i \neq b_i)}{\Sigma_{i=1}^{n}(a_i \cup b_i \neq \emptyset)}.$$

For a given example of the transformable interaction of a pair of words, the normalized estimates of similarity–difference have the form:

$$S(a,b) = \frac{6}{11} = 0{,}55.$$
$$D(a,b) = \frac{5}{11} = 0{,}45.$$

A more complex construction of normalized similarity–difference is determined not by equality, but by the belonging of one coordinate of a vector–word to another coordinate of the second vector, if the coordinates are represented by some sets. In this case, the formulas for calculating the estimates will look like:

$$S(a,b) = \frac{\Sigma_{i=1}^{n}\left(a_i \underset{i=1,n}{\cap} b_i\right)}{\Sigma_{i=1}^{n}\left(a_i \underset{i=1,n}{\cup} b_i\right)}; \quad D(a,b) = \frac{\Sigma_{i=1}^{n}\left(a_i \underset{i=1,n}{\Delta} b_i\right)}{\Sigma_{i=1}^{n}\left(a_i \underset{i=1,n}{\cup} b_i\right)}.$$

Here, set-theoretic operations for the Cantor alphabet will be useful, for example, which are defined by the following quadratic truth tables:

∩	0	1	X	∅
0	1	0	0,5	0
1	0	1	0,5	0
X	0,5	0,5	1	0
∅	0	0	0	0

∪	0	1	X	∅
0	1	1	1	0,5
1	1	1	1	0,5
X	1	1	1	1
∅	0,5	0,5	1	0

Δ	0	1	X	∅
0	0	1	0,5	0,5
1	1	0	0,5	0,5
X	0,5	0,5	0	1
∅	0,5	0,5	1	0

Elementary tables make it possible to convert set-theoretic operations to their norms, the addition of which forms accurate estimates of similarity-difference. For example, for the following two multivalued vectors a=1XXX10X1, b=01X00XX1, the similarity–difference estimates obtained from the numerical truth tables have the form:

$$S(a,b) = \frac{\Sigma_{i=1}^{n}\left(1XXX10X1 \cap 01X00XX1 = 0+\frac{1}{2}+1+\frac{1}{2}+0+\frac{1}{2}+1+1\right)}{\Sigma_{i=1}^{n}(1XXX10X1 \cup 01X00XX1 = 1+1+1+1+1+1+1+1)} = 0{,}56;$$
$$D(a,b) = \frac{\Sigma_{i=1}^{n}\left(1XXX10X1 \,\Delta\, 01X00XX1 = 1+\frac{1}{2}+0+\frac{1}{2}+1+\frac{1}{2}+0+0\right)}{\Sigma_{i=1}^{n}(1XXX10X1 \cup 01X00XX1 = 1+1+1+1+1+1+1+1)} = 0{,}44.$$

The general structure of determining the similarity–difference of a pair of vectors in three vector parallel numerical operations (∩, Δ, ∪) is presented in the following table:

a	1	X	X	X	1	0	X	1
b	0	1	X	0	0	X	X	1
∩	0	0,5	1	0,5	0	0,5	1	1
Δ	1	0,5	0	0,5	1	0,5	0	0
∪	1	1	1	1	1	1	1	1

Thus, two similarity estimates are obtained, which are mutually complementary to each other up to 1: $S(a,b) = 0{,}56$; $D(a,b) = 0{,}44$. The coordinates of the following vectors of intermediate computations are also mutually complementary to 1, which is a condition for validating the process of determining the similarity-difference:

∩	0	0,5	1	0,5	0	0,5	1	1
Δ	1	0,5	0	0,5	1	0,5	0	0

The computational complexity of the algorithm for the synthesis of structural unified metric for transforming one word into another is $Q = (m \times n)^2$.

IV. Frequency-Vector Model and Method for Calculating Similarity

The multivalued structure of a vector pair corresponding to a set of primitive words (T_i, T_j), can be used to effectively determine the similarity of text fragments, and also to calculate the level of plagiarism. A simplified diagram of solving this problem using three vector logical operations can be presented in the form of Fig. 4.

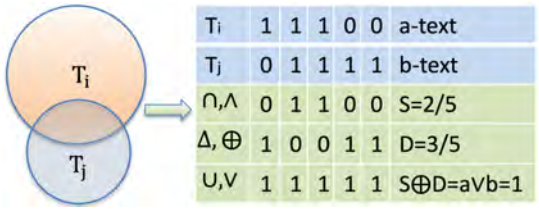


Fig. 4. Diagram for defining similarity of text fragments

In coordinates, words or any other data can be used in vector-sets (T_i, T_j) instead of unit values. However, preliminary unitary coding of words or sentences greatly facilitates the implementation of the algorithm for determining the similarity-difference. It should be noted that instead of a binary code, the coordinates of a vector can be marked by the frequency of words or data in the form of real or integer numbers, and also by the time or other parameters of the components, which does not change the essence of the algorithm for the metric normalized estimation of the similarity of text fragments.

Characteristics of the metric are uniquely determined by essential variables for vectors-sets interacting with each other, does not have variables that are not essential for interacting sets, take into account the frequency of occurrence of each component to calculate the similarity-difference, and also they are a universal model for determining the similarity-difference of any discrete processes and phenomena.

The frequency-vector model of two interacting subsets to determine the similarity-difference is processed according to the following formulas:

$$S(a,b) = \frac{\Sigma_{i=1}^{n}\left(a_i \underset{i=1,n}{\cap} b_i\right)}{\Sigma_{i=1}^{n}\left(a_i \underset{i=1,n}{\cup} b_i\right)} \approx \frac{\Sigma_{i=1}^{n}\left(a_i \min\limits_{i=1,n} b_i\right)}{\Sigma_{i=1}^{n}\left(a_i \max\limits_{i=1,n} b_i\right)};$$
$$D(a,b) = \frac{\Sigma_{i=1}^{n}\left(a_i \underset{i=1,n}{\Delta} b_i\right)}{\Sigma_{i=1}^{n}\left(a_i \underset{i=1,n}{\cup} b_i\right)} \approx \frac{\Sigma_{i=1}^{n}\left|a_i \underset{i=1,n}{-} b_i\right|}{\Sigma_{i=1}^{n}\left(a_i \max\limits_{i=1,n} b_i\right)}.$$

Here, the powers of subsets-frequencies of each component are used, which significantly increase the adequacy of the similarity-difference of texts, processes and phenomena.

Such a modification of the formulas processes not only frequency, but also pure binary representation of vectors-sets. Three operations are used: selection of the minimum value of two coordinates of the same name ($a_i \min\limits_{i=1,n} b_i$), selection of the maximum value of a pair of coordinates ($a_i \max\limits_{i=1,n} b_i$), calculating the absolute value of the difference between two coordinate values $\left|a_i \underset{i=1,n}{-} b_i\right|$. However, arithmetic operations have a drawback here: parallel calculations with the coordinates of vectors cannot be used.

The advantages of the proposed similarity metric of processes and phenomena: 1) The invariance of the frequency-multiple representation of the primitives-data in comparison with the tuple-focused Levenstein distance makes it possible to reduce the computational complexity of the algorithm for determining the similarity-difference from exponential to quadratic. This advantage allows adequate assessing plagiarism of texts in Slavic languages, where it is allowed to change the order of words in sentences. 2) The vector, hardware-oriented model for unitary coding sets of primitives-words makes it possible to calculate the similarity-difference in one automaton cycle. 3) The synthesized unique metric also shows the path of transforming one text into another, as well as the computational complexity of such a transformation, which is defined by differences of vectors-sets. 4) The frequency-vector structure is a universal model for determining the similarity-difference of any discrete processes and phenomena for solving problems of transforming one object into another, making decisions, fault detection, classifying and clustering data. 5) The route of repairing (correction) a faulty product, digital system, software application into a fault-free one based on determining the differences between two metrics. 6) The route of transformation of the destructive genome of the virus into a useful protein based on the determination of the differences in two metrics or the production of antibodies that neutralize the destructive genomes of viruses.

V. Application for Determining the Similarity-Differences of Two Objects

The implementation of the similarity-difference software module is presented in the C++ code. The input–output interface of the software module for calculating the similarity-difference between objects (texts, vectors, matrices, structures), where files-sources, local and integral estimates of similarity-differences, and also control buttons are used, is represented in Fig. 5.

The module has been tested on various text files, including the following pairs: 1) Works of famous authors. 2) Scientific publications of scientists. 3) CVs of experts and newcomers in computing. 4) List of Scopus-publications of individual scientists and researchers. Table 1 of metric vector-set comparison of pairs of text objects and the graph shown in Fig. 6 represent a stable relationship between the estimates obtained by using the software Similarity-module and the software product Unicheck for determining the similarity (plagiarism) [https://corp.eu.unicheck.com/dashboard/library/browser#100071849].

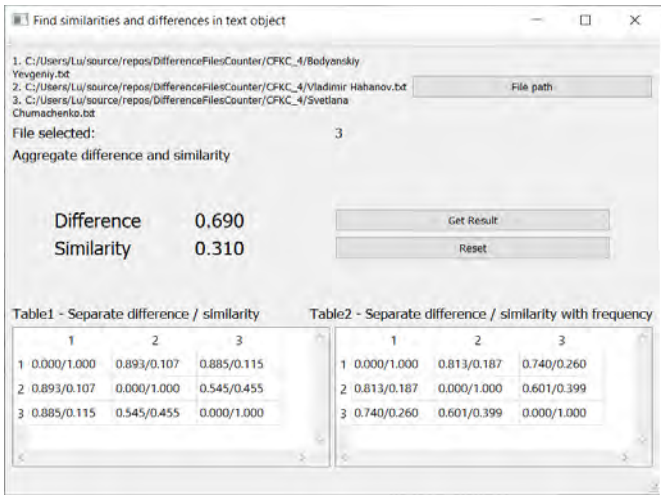


Fig. 5. Interface of C++ application

TABLE 1 – METRIC VECTOR-SET COMPARISON OF PAIRS OF TEXT OBJECTS

Sim-matrix	Resume	Text	Data	Screens	Scopus	Courses	References	E-mails	Papers
Object A (k)	450	150	560	500	40	140	15	3	470
Object B (k)	780	230	120	700	30	100	20	5	500
Similarity U	36	49	47	78	34	57	37	45	70
Similarity N	41	49	42	73	39	52	37	50	64

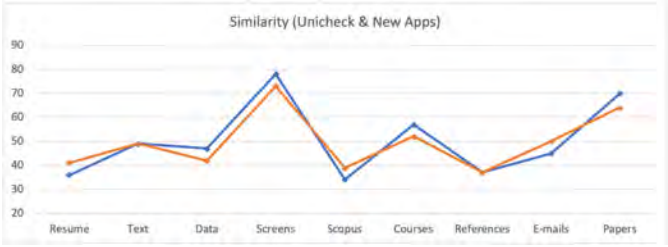


Fig. 6. Correlation between estimates of two software products

Here there is a correlation between the estimates with a maximum deviation of 6 points obtained in various software products, which indicates the consistency of the proposed set-theoretic method for determining similarity-difference of processes and phenomena in order to make adequate decisions.

VI. Conclusion

The structure of cyber-physical computing focused on metric personnel management and decision-making based on comprehensive data collection and subsequent comparison with reference solutions is shown in Fig. 7.

1) The concepts of architectures, models, methods and algorithms of cyber-social computing are proposed, which are relevant for processing big data, decision-making and managing critical systems.

2) The tasks and objective function are formulated to minimize losses associated with failures and repair of critical systems through the development and maintenance of computing structures for metric online decision-making on digital management of critical processes and phenomena based on exhaustive accurate monitoring, use of smart infrastructure

and selection qualified employees meeting reference competencies in education, experience and skills.

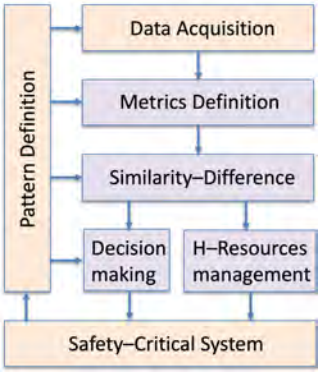


Fig. 7. Management computing structure

3) A frequency set-theoretic method for data retrieval by calculating the similarity-difference of text fragments-objects is proposed, which makes it possible to determine the similarity of objects, the strategy of transforming one object into another, as well as to identify the level of common interests, conflict, plagiarism.

4) C++ code of the frequency set-theoretic method for calculating the similarity-difference of various text fragments-objects is implemented and tested. A comparative analysis of the proposed software application and the existing system for detecting plagiarism is carried out. The results have a high level of correlation for various test examples and low quadratic level of computational complexity of the proposed method.

5) Social similarity and difference. Online bringing together people and countries leads to their on-site division and disintegration of states. The similarity of lifestyles, cultures and territories gives rise to their differences. The opposite is also true, the further the structures, cultures and territories, the closer they are to each other. In closeness, separation is born, and in separation – closeness. Closeness is the cause of many diseases, and separation leads to recovery. Similarity and difference cannot exist without each other and complement each other: S+D=a∪b=1. The further, the closer. The more different the pairs of objects, processes, phenomena or subjects, the closer they are. A harmonic is the genome of the development of nature and society. The phase of absolute similarity gives rise to the process of differentiation. The opposite is also true. The prosperity phase of statehood is the beginning of the process of its degradation and disintegration.

REFERENCES

[1] A. Drozd, V. Kharchenko, S. Antoshchuk et. al., “Checkability of the digital components in safety-critical systems: problems and solutions,” IEEE East-West Design & Test Symposium, Sevastopol, Ukraine, 2011, pp. 411–416.

[2] O. Drozd, V. Kharchenko, A. Rucinski et. al., “Development of Models in Resilient Computing,” IEEE International Conference DESSERT, Leeds, UK, 2019, pp. 2-7.

[3] V. Hahanov, E. Litvinova, and S. Chumachenko, “Green Cyber-Physical Computing as Sustainable Development Model,” In the Book “Green IT Engineering: Components, Networks and Systems Implementation”. Editors V. Kharchenko, Y. Kondratenko, J. Kacprzyk, Springer, 2017.

[4] D. Tarraf, “Control of Cyber-Physical Systems,” Workshop held at Johns Hopkins University, March 2013, Springer, 2013.

[5] R. Guo, G. Mao, Y. Liu, Y. Liu, J. Wang, R. Cui, "The Method of Similarity-Difference Comprehensive Evaluation on Test Paper Quality in Colleges and Universities and Its Application," 2009 Second International Conference on Education Technology and Training, Sanya, 2009, pp. 227-230.

[6] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, L. Zheng, "Deep Hybrid Similarity Learning for Person Re-Identification," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 11, Nov. 2018, pp. 3183-3193.

[7] T. Komori, Y. Hijikata, T. Tominaga, S. Yoshida, N. Sakata and K. Harada, "Real Friendship and Virtual Friendship: Differences in Similarity of Contents/People and Proposal of Classification Models on SNS," 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, 2018, pp. 354-360.

[8] K. Lin, "New Vague Set Based Similarity Measure for Pattern Recognition," 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Toyama, Japan, 2019, pp. 15-21.

[9] V. Hahanov, S. Chumachenko, E. Litvinova, and M. Liubarskyi, “Qubit Description of the Functions and Structures for Computing ,” Proc. of IEEE East-West Design and Test Symposium, Yerevan, 14-17 Oct., 2016. pp. 88-93.

Generaled Installation of Fuzzy Linear Automaton

Dmitriy V. Speranskiy
Department of "Control Systems of Transport Infrastructure"
Russian University of Transport (MIIT),
Moscow, Russia
Speranskiy.dv@gmail.com

***Abstract*— The article considers the linear automata whose fuzziness is embedded in the equations of their functioning. The method is proposed for constructing generalized sequences that lead them to a certain set of known initial states. The concepts of optimality sequences by various criteria are introduced and the method of their construction is proposed. This problem is very relevant for the diagnosis of digital systems.**

Keywords— ***fuzzy linear automaton, generalized installation sequence, optimal installation sequences, synthesis method.***

I. INTRODUCTION

The deterministic finite automat [1] is a convenient and adequate model for many real technical systems, including various digital devices (DD). However, in the process of model construction, the information about their functioning is sometimes approximate (fuzzy). This is explained by the fact that such information is sometimes formulated in terms of poorly formalizable concepts, for example, "almost equal", "close enough", "many", "few", etc. Presentation of such information in the traditional mathematical language is difficult and leads to coarsening of the model. Obviously, appropriate tools are required to adequately reflect the fuzziness of such information. An important step towards the creation of suitable tools was the theory of fuzzy sets, developed by l. Zade [2]. The proof of usefulness and effectiveness of the concept of fuzziness has been confirmed in numerous applications in many subject areas, partly reflected, for example, in [3].

The fuzziness in the description of the system models may occur both to the algorithms of their functioning and to the initial data, as well as to their input and output signals. Various types of fuzzy automata (FA) became the subject of research soon after the fundamental article by L. Zadeh [2]. The monograph by D. Dubois and G. Prada [4] traces the evolution and provides a bibliography of publications in this direction. Without going into details, we note that the investigated varieties of FA differed from each other in different requirements for the initial state (clear and fuzzy), for the functions of transitions and outputs (clear and fuzzy), for the functions of belonging of FA reactions at different input signals, etc.

The article simulates the possibility of the appearance of fuzziness due to the use of a special method for describing the algorithms of the system's functioning, and fuzzy automata (FA) are used as its model.

A fragment of the theory of experiments with fuzzy automata, in general case nonlinear, was developed in [5]. The proposed article is related to the same subject, but the object of research is the fuzzy linear automaton (FLA). Interest to linear models is caused by their extensive use both in theory (for example, synthesis of special counters, encoders) and in practical applications (error detection in codes, message encryption, etc.) [6]. In addition, linear automata have specific features that allow them to obtain some results (e.g. conditions of existence of different types of experiments) in a much easier and more convenient form for calculation and verification than for automata of general type. The above factors justify the interest in the theory of experiments with fuzzy automata.

The purpose of this article is to investigate the problem of a generalized installation of FLA, which is used, among other things, as a model of the DD. The term "installation" is used here in one of the generally accepted interpretations - as preparatory measures before the introduction of the equipment into the functioning process. For example, before a DD can be diagnosed with a test, it must be converted to the specified number of known conditions. The FA installation can be performed using two types of input sequences - homing and/or synchronizing. The use of the homing sequence implies mandatory monitoring of the DD outputs, while the use of a synchronizing sequence to observe the output values is not required. Further, both types of sequences will be referred to the common term - installation (IS).

The method of their synthesis described in the article generates as a result (according to the terminology [1,7]) the socalled homing sequences, but if FLA there is synchronized FA then this will be the construction of a synchronizing sequence for FA.

The solution to the problem of a generalized DD installation can be useful for some practical applications.

Let's mention two of them. The first relates to the problem of technical diagnostics of DD, which, as we know, requires DD testing. However, before testing, the device must be set to a known initial state (or to a small number of known states), because only then it will be possible to predict the behavior of the DD.

The second is the problem of safe functioning of complex systems based on the synchronization of their components interaction. For example, for air transport systems (ATS) this is related to the distribution of aircraft flow near the airport. Sometimes the airport is so heavily loaded that take-off and landing on lanes occur every minute and the dispatcher has to make quick decisions that ensure flight safety. In particular, the dispatcher has to distribute their flow to different echelons, preventing dangerous rapprochement of the boards with each other. Such requirement in the ATS is provided by the dispatcher by giving a command to the pilot to change the echelon. The solution of this problem, based on the application of the automatic model, is proposed in [8].

The same article gives an example of the solution of a specific problem of the above type, reducing it to the construction of the socalled generalized synchronizing sequences, which were first introduced in [7]. In [8] it is argued that

such a distribution of boards by echelons is used for airport personnel training.

II. DESCRIPTION OF THE FLA MODEL

The model of the fuzzy automaton used below was introduced in [5]. Recall that in [5] the fuzzy finite automaton (FA) is called the five of objects.

$$A = (S, X, Y, \delta, \lambda), \quad (1)$$

where $S = \{s_1, ..., s_n\}$ - finite set of states, $X = \{x_1, ..., x_m\}$ - finite set of inputs, $Y = \{y_1, ..., y_p\}$ - finite set of outputs, $\delta: S \times X \times [0, 1] \to S$ - transition function, $\lambda: S \times X \times [0,1] \to Y$ - output function. The function δ has the following interpretation: FA, which is in the state *s*, when input *x* is applied, returns to the state s', and the value of the function of the set of the elements (s, x, s') in the fuzzy subset is equal to some value $q \in [0,1]$.

The function δ introduced above actually generates many fuzzy matrices for each input $x \in X$:

$$T(x) = [\mu_x(s_i, s_j)], \ 1 \le i, \ j \le n. \quad (2)$$

Here the value $\mu_x(s_i, s_j) \in [0,1]$ is an estimation of the degree of possibility of transition from state s_i of FA to the state s_j when the input *x* is applied.

The function of the output $\lambda: S \times X \times [0,1] \to Y$ has the following interpretation: FA, which is in the state *s*, when input *x* is applied, initiates the output *y*, and the value of the set of elements (s, x, y) to the fuzzy subset $S \times X \times Y$ is equal to some value $q \in [0,1]$.

The function actually generates a set of matrices $\Lambda(x)$ for each input $x \in X$:

$$\nu_x(s_i, y_j) \in [0, 1], \ 1 \le i \le n, \ 1 \le j \le p. \quad (3)$$

Here, the value $\nu_x(s_i, y_j) \in [0, 1]$ is an estimation of the degree of possibility of FA output y_j at its transition from the state s_i when the input *x* is supplied.

Before formulation of the definition of FLA, which is the object of our study, let us recall what is the "classical" deterministic linear automaton (LA) [6,7].

LA is a system with a finite number l of input poles and with the finite number m of output poles. It is assumed that the input signals take values from the field *GF*(*p*) = {0, 1, .., *p*-1}, where *p* is a prime number. The LA state is understood as an ordered set of delay element states (let us denote their number in *n*), which are part of the LA. The number *n* is usually called the LA dimension.

The LA functioning above the field *GF*(*p*) is defined by the following transitions and outputs equations:

$$\bar{s}(t+1) = A\,\bar{s}(t) + B\,\bar{u}(t), \quad (4)$$
$$\bar{y}(t) = C\,\bar{s}(t) + D\,\bar{u}(t). \quad (5)$$

Here t is the moment of discrete time; $A = [a_{ij}]_{n \times n}, B = [b_{ij}]_{n \times l}, C = [c_{ij}]_{m \times n}, D = [d_{ij}]_{m \times l}$ are characteristic matrix. Elements of all these matrices are the elements of the *GF*(*p*) field.

The input vector $\bar{u}(t)$, the output vector $\bar{y}(t)$ and the vector-state $\bar{s}(t)$ are ordered sets of columns of elements of the same field:

$$\bar{u}(t) = [u_1(t), ..., u_l(t)]',$$
$$\bar{y}(t) = [y_1(t), \ldots, y_m(t)]',$$
$$\bar{s}(t) = [s_1(t), \ldots, s_n(t)]'.$$

Now let's move on to the description of the FLA. In principle, the uncertainty of FLA functioning can be implemented in different ways. Thus, this fuzziness can be put separately either in each matrix of equations (4) and (5), or simultaneously in several of these matrices. It can be shown that the choice of one of the listed fuzzy methods does not fundamentally differ from the other due to the possibility of converging one method to another. Therefore, let us choose the one that will simplify the subsequent presentation and calculations.

Next, let us agree on the fuzziness of FLA functioning to model, for example, using only matrix *B*, considering the other matrices constant. Elements of matrix *B* will be written in the form $b_{ij} = b_1 \vee b_2 \vee \ldots \vee b_f$, where b_i $(i = 1, ..., f)$ - elements from the field *GF*(*p*) above which the FLA is given. This recording means that at any machine tact of automata time the element can be equal to any of the elements b_i.

Thus, the external form of recording the equations of transitions (4) and outputs (5) of FLA does not change.

III. INSTALLATION PROBLEM

Let us first recall the concepts of synchronizing (SS) and homing sequences (HS) for a deterministic finite automat [1,7]. The input sequence *p* is called a SS, if its supplied transfers the automaton to the same final state regardless of the state from which it started. This definition assumes that a set of S_0 possible (acceptable) initial states is known in advance. Note, that this set may coincide with the whole set *S* of states of the automaton. In [5], we introduced a similar concept for FA, which naturally also applies to FLA: the input sequence $p = \bar{u}(0), \bar{u}(1), ..., \bar{u}(k)$ is called the generalized SS (GSS) for FA (FLA) if

$$\delta\ (S_0, p) | < | S_0 |. \quad (6)$$

Here $\delta(S_0, p)$ is a set of all states to which transfers FLA from the states of the set S_0 when supplied the sequence *p*. An input sequence $p = \bar{u}(0), \bar{u}(1), ..., \bar{u}(k)$ is called a generalized HS (GHS) for FA (FLA) if

$$\forall_{s_1, s_2 \in S_0} \lambda(s_1, p) = \lambda(s_2, p) \to | \delta(S_0, p) | < | S_0 |. \quad (7)$$

The meaning of using both types of installation sequences (IS) is to reduce the power (fuzziness), into which the FA (FLA) goes when it is applied. We will further call them as Final Installation Sets (FIS). It is clear that the definition of GSS and GHS coincides with the definition of SS and HS from [1,7] for a deterministic automaton, if $| \delta(S_0, p) | = 1$.

Based on the inequalities (6) and (7), we can talk about the different degree of fuzziness of FLA sets $\delta(S_0, p)$. In general, it is natural to assume that the higher the installation quality imply, the less fuzzy. We will consider $|\delta(S_0, p)|$ as an indicator of this quality and call the indicator of fuzziness. Ideal is the case when $|\delta(S_0, p)| = 1$.

Another criterion by which we can compare IS is their length. Obviously, of two IS with the same $|\delta(S_0, p)|$, the one that is shorter in length is considered the best.

In [5] it is noted that the transition of FA from a state *s* to a certain state $s' \in \delta(S_0, p)$ when IS is *p* is possible by different trajectories (paths). Among all such possible trajectories, let there be a trajectory $\gamma = s \to s_{t_1} \to s_{t_2} \to ... \to s_{t_k} \to s'$. Let's match it with a number $M(p, \gamma)$ that is called the degree of transition along the trajectory γ. For expression $M(p, \gamma)$ through the transitions function described in [5], one can use one of the following options suggested in a number of publications:

$$M_1(p, \gamma) = \min[\delta_\mu(s, p_{t_1}, s_{t_1}), \delta_\mu(s_{t_1}, p_{t_2}, s_{t_2}), ..., \delta_\mu(s_{t_k}, p_{t_k}, s')], \quad (8)$$

$$M_2(p, \gamma) = \max[\delta_\mu(s, p_{t_1}, s_{t_1}), \delta_\mu(s_{t_1}, p_{t_2}, s_{t_2}), ..., \delta_\mu(s_{t_k}, p_{t_k}, s')]. \quad (9)$$

Let p there be some IS, and q_1, q_2 - two different trajectories of transitions FA at IS supplying. Let $M_i(p, q_1), M_i(p, q_2), i = 1, 2$ is the degrees of transitions along the trajectories q_1, q_2. Using the above values, one can enter the preference ratio between the trajectories q_1 and q_2.

To sum up, it should be noted that in general, IS for fuzzy automata can be compared by three different criteria:

1. by the power of the FIS that generates the generalized IS;
2. by the length of the generalized IS;
3. by estimating the degree of transitions of trajectories (characterized by M_1 and M_2 values) selected (by maximum or minimum) from a set of all trajectories corresponding to the comparable IS.

Below we solve the problems of constructing such IS for FLA which are optimal by one or more given criteria.

IV. SYNTHESIS OF GENERALIZED IS GENERATING OPTIMAL POWER FINAL INSTALLATION SETS

At the beginning we will consider the task of constructing IS that minimize the power of the FIS for a given FLA and a given set S_0 of its initial states. We will call such IS optimal.

Let's assume that for a given FLA $\widehat{A}$ generalized IS exist, and the set of all such IS we will denote through $R(\widehat{A})$. Let's find all optimal IS. In general, they can be more than one.

Let's $p = \overline{u}(0), \overline{u}(1), ..., \overline{u}(k)$ there be some optimal IS.

A set of all states (FIS) initiated by IS is a set of vector-columns of the form $(s_1, s_2, ..., s_n)'$, where *n* is the FLA dimension. We start solving the problem of finding the optimal IS by constructing a graph, which we call an installation (IG).

It is further assumed that the reader is familiar with classical notions and terminology used in [1] when describing the successor tree of the finite automaton.

It should be noted that the rules of constructing the IG are in many respects similar to the rules of constructing the successor tree for the deterministic automaton in [1].

The IG consists of vertices placed on successive levels. The zero level contains one vertex (node) S_0, which consists of a set of all allowed initial states of FLA $\widehat{A}$. This set is further treated as an identifier (name) of the corresponding vertex. Different subsets of the *S* set of FLA states, localized in the nodes of the other levels of the IG, are interpreted in the same way. If *X* is the input alphabet of FLA, where $|X| = m$, then each node of level k of the IG ($k = 0.1, ...$) has *m* outgoing edges, each of which corresponds to one of the symbols of the input alphabet of FLA.

The nodes $S_0^1(x)$ of the first level is a subset of the states of FLA, into which it passes from the states of the set S_0 when input symbol *x* is fed to the automaton.

The elements of the set $S_0{}^1(x)$ are calculated by formula (4). At the same time, matrix *B* in (4), according to the method of introducing fuzziness in the functioning of FLF previously agreed by us, appears to be some set of *z* pairs of different matrices of dimensional $n \times l$, (same as in matrix *B*). Each of them is one of the possible variants of matrix *B* obtained from a specific distribution of alternative methods of introduce fuzziness. In other words, in calculating the elements of the set $S_0^1(x)$ in (4), *z* pairwise pairs of different matrices replacing *B* will be used.

Let us present all the mentioned possible variants as a set of $B = \{B_1, B_2, ..., B_z\}$. Let's assume that on every tact of the automaton time during the process of feed to the input of the sequence there may be any of the matrices $B_1, B_2, ..., B_z$ as matrix *B*. Thus, the set $S_0^1(x)$ will be a union of *z* subsets of the set of FLA states into which it is possible to transitions from the states of the set due to the fuzziness in matrix *B*. Similarly, other nodes in the IG (subsets of the FLA state set) are constructed and named in the same way.

Let us describe the differences between the calculation of the successors of A-group *G* in [1] by some input signal for the successor tree and the similar procedure for the IG. Recall that A-group *G* is composed of - σ sets of states in which there can be repetitive elements. IG group *G* is a classical set (there are no identical elements in it) and its successor is also a classical set. In other words, in the construction of the IG, the set is obviously transformed into a regular set.

For example, if there is A-group in [1] $G = (\{2, 3, 4, 4, 2\}, \{5, 2, 5\})$, it is converted into a IG into two nodes, whose names are subsets {2,3,4} and {5,2}.

It follows that if in the installation tree of [1] a node of zero level (some group *G*) contains m states, then all its subsequent successors at any subsequent level also contain by *m* states, but this fact has no place in the construction of IG. In our example, group $G = (\{2, 3, 4, 4, 2\}, \{5, 2, 5\})$ contained 8 states and its successor in [1] will also contain 8 states. However, when constructing the IG, two vertices will be obtained, the total number of states of which is 5.

As well as in [1] when constructing the successor states in a IG by some input signal, each subset of g from group G is divided into subsets such that two states from g are included in the same subset if and only if they produce the same outputs per input signal. All such states are enclosed in curly brackets.

It is clear that during the construction of the IG, from each node will come the edges marked with all input signals of FLA. If the successor node "received" a new name, which is absent in the IG that was construct up to this point, it is added to the IG, and from the generating node an edge is held marked with a corresponding input symbol is drawn into it. If a node already exists in the IG, the edge marked with the appropriate input symbol is drawn from the originating node into the IG. Obviously, the process of constructing a IG for an FLA is always finite.

Let's illustrate the construction of the IG on the example of FAL above the field $GF(2)$, which has a set of acceptable initial states $S_0 = \{0,1,2,3\}$, and the characteristic matrices in equations (4) and (5) have the following form:

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \vee 1 \\ 1 \\ 0 \vee 1 \end{bmatrix}, \; C = [0\,1\,0], D = [1].$$

Because for it $n = 3$, $l = 1$ it has $2^3 = 8$ states and an input alphabet $X = \{0,1\}$. Let us agree to reduce the notation to the states of this automaton to denote by numbers

$$0 = (000)', 1 = (001)', 2 = (010)', 3 = (011)', 4 = (100)',$$
$$5 = (101)', 6 = (110)', 7 = (111)'.$$

Since in the given matrix B the elements b_{11} and b_{13} have two variants of elements to choose from, then further we will consider, for example, the following variants of fuzziness distribution in it:

$$B = \{B_1, B_2, B_3\} = \left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}.$$

Further, to simplify the presentation, we will agree to consider that the values of the variables of the possibility to choose any variant of the matrix from the set B are equal among themselves. The calculation $\delta(0,0)$ and $\delta(0,1)$, i.e., the states, to which FLA passes from the state 0 when input signals 0 and 1 are given, will be performed according to formula (4):

$$\overline{s}(1) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \end{bmatrix} [0],$$

$$\tilde{s}(1) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \end{bmatrix} [1].$$

In them, matrix B can be any of the matrices B_i, $i = 1,2,3$. As a result, we get that for any variant B_i,

$$\delta(0.1) = \left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\} = \{2,6,7\}.$$

Other transitions of FLA are calculated in the same way.

$$\delta(1,0) = \{0\}, \delta(1,1) = \{2,6,7\}, \delta(2,0) = \{4\},$$
$$\delta(2,1) = \{2,3,6\}, \delta(3,0) = \{4\}, \delta(3,1) = \{2,3,6\}$$

Further, the calculation $\lambda(0,0)$. $\lambda(0,1)$, i.e. the outputs of FLA from the state 0 when input signals 0 and 1 are supplied, will be performed by formula (5):

$$\lambda(0,0) = [0\,1\,0] \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + [1] \cdot 0, \quad \lambda(0,1) = [0\,1\,0] \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + [1] \cdot 1.$$

Since in (5) matrices C and D are constant, the values of its outputs $\lambda(s,x)$ for any states s are also constant with the any method we have chosen.

Continuing this process further for the rest of the FLA states for all input signals of the FLA under consideration, let us build its complete transient/output table below:

State	Input	
	0	1
0	{0/0}	{2/1,6/1,7/1}
1	{0/0}	{2/1,6/1,7/1}
2	{4/1}	{2/0,6/0,3/0}
3	{4/1}	{6/0,2/0,3/0}
4	{5/0}	{7/1,3/1,2/1}
5	{5/0}	{7/1,3/1,2/1}

Now let's get back to constructing an IG for our FLA example. It follows from the table of transitions that from the initial vertex {0, 1, 2, 3} the edge marked by the symbol $x = 0$ should lead to the node with the name {0, 4}, because $\delta(0,0) \cup \delta(1,0) \cup \delta(2,0) \cup \delta(3,0) = \{0,4\}$, and the edge marked by the input $x = 1$ - to the node {2, 3, 6, 7}, because $\delta(0,1) \cup \delta(1,1) \cup \delta(2,1) \cup \delta(3,1) = \{2,3,6,7\}$. Continuing this process, we will obtain for the considered FLA and the given set S_0 of its allowable initial states the corresponding IG shown in Fig.1.

From this graph we can see that all its nodes of the 1-st and 2-nd levels (more precisely, subsets of FLA states that are their names), except for one with the name {2,3,6,7}, satisfy the inequality (7). This means that all input sequences corresponding to the paths in IG from the initial vertex to each of them are IS. In the IG in Fig.1 there are three vertices ({0.4}, {1,4}, {0,5}) with the fuzziness index 2 and two vertices ({2,6,3},{2,6,7}) with the index 3. Thus, three different optimal IS (with index 2) can be constructed for the considered FLA example: p = 0; p = 0,0; p = 1,0. If we evaluate the IS by length, then the first of the given optimum is the best by this criterion too.

Let us make one add-on concerning the peculiarity of constructing a IG for an FLA having a synchronizing IS. It should be noted that this fact can be easily checked. The corresponding necessary and sufficient condition established

in [7] for deterministic automata turns out to be true also for FLA due to the assumption made about the mechanism of fuzziness in their functioning. It is proved that this condition consists in the nilpotence of the main characteristic matrix A of the automaton, i.e., in the existence of such a whole k that A^k there is a zero matrix. In this case, for FLA, the construction of the IG can be simplified.

When constructing the IG, in case if FLA has SS, there is no need to observe its outputs. It follows that now there is no need to divide up the set of successor states into subsets with the same outputs. Consequently, the number of nodes in the IG can generally decrease in this situation. Thus, if the FLA considered in the above example had been SS, then in the in Fig. 1 the nodes {2,6,7} and {2,6,3} should be combined into one node with the name "2,3,6,7". Thus, for the synchronized FLA the installation graph will be more compact.

V. SYNTHESIS OF IS THAT ARE OPTIMAL IN LENGTH AND ESTIMATION FLA TRAJECTORIES

Let us now consider the task of finding IS with a better estimate of the FLA trajectory that goes it, for example, in FIS {1,4}. As can be seen from Fig.1, all these trajectories are the following simple paths (the simplicity of the path is understood as it is customary in the theory of graphs) through the IG from the vertex S_0 to the node {1,4}:

1. $\{0,1,2,3\} \xrightarrow{1} \{2,6,7\} \xrightarrow{0} \{1,4\};$
2. $\{0,1,2,3\} \xrightarrow{1} \{2,6,3\} \xrightarrow{0} \{1,4\};$
3. $\{0,1,2,3\} \xrightarrow{0} \{0,4\} \xrightarrow{1} \{2,3,6,7\} \xrightarrow{0} \{1,4\};$
4. $\{0,1,2,3\} \xrightarrow{0} \{0,4\} \xrightarrow{0} \{0,5\} \xrightarrow{1} \xrightarrow{1} \{2,3,6,7\} \xrightarrow{0} \{1,4\}.$

Using the method described above, we will match each trajectory with the numbers i = 1,2,3,4 numbers $M_1(p,i)$ and $M_2(p,i)$, where p - the input sequences that generate estimates of the degree of transitions along the i-th trajectory. To calculate the estimations, it is necessary to know the matrices of $T(x)$ type described above for each input x.

It should only be noted that the elements of these matricesare the numbers from the interval [0, 1], which are expert estimations of the degree of transitions from one state of FLA to another. These matrices themselves are constructed on the basis of the table of transitions of FLA. The choice of the best trajectory is made on the basis of estimations obtained for all possible trajectories of FLA corresponding to the IS under study.

It follows from the above that there is a fundamental difference between the installation problem for a deterministic automaton and a similar problem for an FLA. For the FLA this problem is in general a multi-criteria one, while for the deterministic automaton only one criterion is always used for the evaluation of different IS – it is length [1].

For solving multi-criteria problems, formalization is necessary, which requires expert evaluations of the criteria themselves and identification of relations between them. These used criteria are known may be either contradict each other, act in one direction or be independent.

At present, a number of approaches to solving multi-criteria problems are used. One of them is related to singling out the most important criterion and searching for a solution to an optimization problem only for this criterion. In this case, other criteria of the problem in question play a role of additional constraints. The second approach is based on ordering the given set of criteria and performing a sequential optimization for each of the criteria separately. The third approach is based on reducing a given set of criteria to one. Typically, this reduction is performed by introducing weighting coefficients for each criterion. The weighting ratios themselves is expert estimates. Much articles have been devoted to the development of the above approaches and methodology of searching for solutions to multi-criteria problems. For example, various aspects of this problem have been studied in detail in [9,10].

VI. CONCLUSION

The presented results show that the installation problem for FLA, in contrast to the same problem for a deterministic automaton, is, on the whole, multi-criteria. The methods proposed in the article make it possible to synthesize the generalized IS which are the best for each of the three criteria separately. These methods are based on the use of a special graph (IG), the method of construction of which is described in the article.

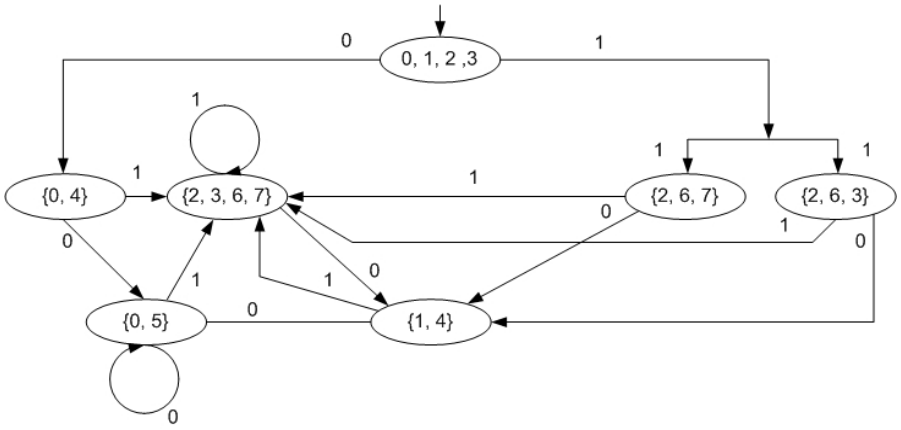


Fig. 1. Installation graph for example of FLA

REFERENCES

[1] A. Gill “Introduction to the theory of finite-state machines”, New York, Mc-Graw-Hill,1962. 218 p.

[2] L.A. Zadeh “Fuzzy sets”, Inf. Control. 1965. № 8. pp.338-353.

[3] T. Terano, K. Asai, and M. Sygeno “Applied fuzzy systems”, Eds. Terano T., Asai K., Sygeno M. Moskow, Mir,1993, 184 p. (in Russian)

[4] D. Dubois, H. Prade “Fuzzy Sets and Systems: Theory and Applications”. N Y: Academy Press, 1980, 236 p.

[5] D.V. Speranskiy “Experiments with fuzzy finite state machines”, Automation and Remote Control. 2015, Vol. 76, №2, 2015, pp. 278-291.

[6] A. Gill “Linear sequential circuits. Analysis, synthesis and applications”. New York, Mc-Graw-Hill book company, 1967, 215 p.

[7] D.V. Speranskiy “Lectures on the theory of experiments with finite automata”, Moscow, Internet-Universitet Informatsionnyx Texnologiy: BINOM. Laboratoriya znaniy, 2010, 287 p. (in Russian)

[8] A.F. Rezchikov, A.S. Bogomolov, V.A. Ivashenko, L.Yu. Filimonyuk “Automation model-based approach to safety of complex systems”, Upravlenie bolshimi sistemami. 2014. Iss. 54. pp.174-194. (in Russian)

[9] R.L. Kini, H. Rayfa “Decision-making under many criteria: preference and substitution”, Moscow, Radio I svyaz, 1981.560 p. (in Russian)

[10] R. Steuer “Multiple criteria optimization: theory, computation and applications”, Moscow, Radio i svyaz, 1992.504 p. (in Russian).

Reception of DPSK-QAM Combined Modulation in Fast Fading Channels by Searching over DPSK Hypotheses

Alexander B. Sergienko
Department of Theoretical Fundamentals of Radio Engineering
Saint-Petersburg Electrotechnical University "LETI", Saint Petersburg, Russia
Email: sandy@ieee.org

***Abstract*—A reception method is presented for the modulation scheme that combines differential phase shift keying (DPSK) and quadrature amplitude modulation (QAM). QAM symbols are inserted between DPSK symbols, and the phase offset of QAM symbols is determined by the preceding DPSK symbol. Such modulation scheme is suitable for noncoherent reception and its bit error rate (BER) performance is better than that of a pure DPSK with the same spectral efficiency. In the case of fast fading, optimal reception of such modulation requires complete search over a large signal set that leads to prohibitive computational cost. The proposed method is a feasible simplified approximation of the generalized likelihood ratio test. The search is performed only over DPSK hypotheses while for QAM symbols hard decisions are used instead of a complete search. The simulation results showed that the proposed method reduces error floor in channels with fast fading by several orders of magnitude. It is also shown that the optimal DPSK constellation size decreases with Doppler spread. Potential BER gain decreases with the number of QAM symbols inserted between DPSK symbols. In the case of one QAM symbol, BER drop factor can be as high as 5.**

***Index Terms*—differential phase shift keying, quadrature amplitude modulation, fading channel, fast fading, channel state estimation, noncoherent reception**

1. Introduction

Standard method for reception of signals that require coherent detection relies on the channel state information (CSI) estimation. To get these estimates, pilot signals are used. They consume power and time-frequency resource, therefore this overhead should be kept as small as possible.

Different approaches are known for reducing the fraction of system resources allocated to the transmission of pilot signals. In [1], [2], blind estimation of CSI from data symbols was used to increase estimation accuracy while dedicated pilot symbols were used primarily to resolve angular ambiguity inherent for blind estimates. The main drawback of this approach is that it requires sufficiently long symbol sequences (3...4 times signal constellation size) for reasonable accuracy of blind CSI estimation. Underlying assumption in this method is that the channel gain is a constant, so its use for fading channels is limited to the case of very slow fading.

To tolerate fast fading, modulation schemes are used that allow noncoherent reception over a small number of symbols. The most used linear modulation of this class is differential phase shift keying (DPSK) where noncoherent decisions can be made from symbol pairs (see, for example, [3, Chapter 5]), but with the increase of the DPSK constellation size noise immunity sharply drops. Further development of this approach led to differential amplitude/phase modulation [4], however, some coding is required for its use, and jointly demodulated signal fragments can not be very short (in [4], sequences of 6 symbols are considered).

Another direction is the use of modulated pilots that allow estimation of the channel gain and at the same time can be used for data transmission. In [5], a modulated pilot signal is proposed for orthogonal frequency division multiplexing (OFDM) systems. This pilot occupies two frequency-adjacent OFDM cells. One cell is modulated using phase shift keying (PSK), it allows to estimate the channel gain magnitude. The second cell is modulated using amplitude shift keying, it allows to estimate the channel gain phase. If the channel gains for these two cells are close to each other, such pair of symbols can be used as a pilot for reception of the other data symbols. As pilot signal in this case is not fixed, this approach requires more complex signal processing than in the case of conventional pilot signals.

In [6], a combination of DPSK and quadrature amplitude modulation (QAM) was proposed to create a signal set suitable for noncoherent reception. It was shown that in additive white Gaussian noise (AWGN) channel this modulation scheme can provide better combinations of spectral and power efficiencies than pure DPSK.

In AWGN channel, computationally simple reception method for this modulation scheme can be based on using demodulated DPSK symbols as pilots for demodulation of QAM symbols. In channels with fast fading, where channel gain notably varies between DPSK symbols, the reception problem is far more difficult. One possible method of reception will be considered in this paper.

The paper has the following structure. In Section 2, signal and channel model is described. In Section 3, reception methods are presented. Simulation results are considered in Section 4. Finally, conclusions are drawn in Section 5.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

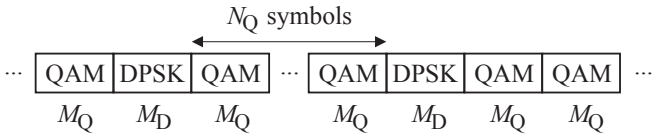


Figure 1. Sequence of DPSK and QAM symbols

2. Signal and Channel Model

The idea of combined DPSK-QAM modulation scheme can be formulated as follows. There is a sequence of M_D-ary DPSK symbols, and every pair of adjacent DPSK symbols is separated by N_Q M_Q-ary QAM symbols (see Figure 1). To make possible reception of QAM symbols without additional pilot signals, DPSK symbols are used as a phase reference, i. e., every DPSK symbol defines the phase offset for N_Q subsequent QAM symbols.

Mathematical description of such signal has the following form:

$$s(t) = \sum_{k=-\infty}^{\infty} \sum_{m=0}^{N_Q} a_{k,m} g(t - (k(N_Q+1)+m)T), \quad (1)$$

where k is the number of the block consisting of $N_Q + 1$ symbols, m is the symbol number inside the block, $g(t)$ is the signal pulse, T is the symbol period, $a_{k,m}$ are modulation symbols having the following form:

$$a_{k,m} = \begin{cases} a_{(k-1),0} \exp\left(j \frac{2\pi}{M_D} d_{k-1}\right), & m = 0, \\ a_{k,0} c_{k,m}, & m \neq 0, \end{cases} \quad (2)$$

where $d_k \in \{0, 1, \ldots, M_D - 1\}$ are integer-valued representations of DPSK symbols, $c_{k,m} \in \mathcal{C}$ are complex-valued QAM symbols from the M_Q-ary QAM constellation $\mathcal{C}$. The average powers of DPSK and QAM symbols are set to unity as power optimization results presented in [6] showed that optimum powers are very close to one. Therefore, power optimization gain is very low, and this optimization will not be considered here.

This scheme can be treated as a transmission of overlapping blocks with the length of $N_Q + 2$ symbol intervals (Figure 2). These blocks are signals from a signal set suitable for noncoherent reception. This signal set consists of $M = M_D M_Q^{N_Q}$ signals and thus conveys $m_D + N_Q m_Q$ data bits, where $m_D = \log_2 M_D$ and $m_Q = \log_2 M_Q$ are the numbers of bits in the DPSK and QAM symbols, respectively. The structure of the signals from this signal set is shown in Figure 3, it can be described as follows:

- one real-valued pilot symbol equal to one;
- N_Q QAM symbols with unit average power;
- one PSK symbol with magnitude equal to one.

Spectral efficiency (SE) of this modulation scheme is

$$\mathrm{SE} = \frac{m_D + N_Q m_Q}{N_Q + 1} \text{ bits per symbol.} \quad (3)$$

Channel model with flat fading and AWGN is adopted. The fading is considered fast enough that we cannot ignore channel gain changes inside signal blocks, but at the same time slow enough to ignore multiplicative distortions of signal pulses.

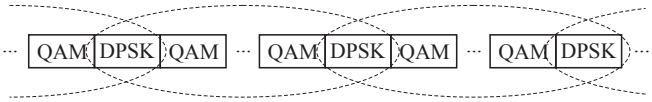


Figure 2. Overlapped signal blocks

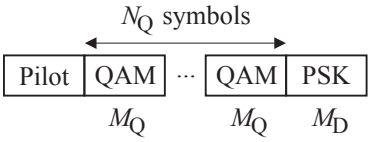


Figure 3. Signal set structure

Received samples after the matched filter (ideal time synchronization is assumed) have the form

$$x_{k,m} = h_{k,m}\, a_{k,m} + n_{k,m}, \quad (4)$$

where $h_{k,m}$ is the channel gain for the m-th symbol of the k-th block, and $n_{k,m}$ are the samples of complex-valued AWGN with average power P_n.

Because of block overlapping, DPSK symbols will be treated as belonging to two blocks simultaneously:

$$\begin{aligned} x_{k,0} &= x_{(k-1),(N_Q+1)}, & h_{k,0} &= h_{(k-1),(N_Q+1)}, \\ a_{k,0} &= a_{(k-1),(N_Q+1)}, & n_{k,0} &= n_{(k-1),(N_Q+1)}. \end{aligned} \quad (5)$$

Rayleigh channel model is assumed, so that $h_{k,m}$ are sample realizations of correlated zero-mean unit-variance complex-valued Gaussian random variables. Correlation between these variables is described by Jakes model [7]:

$$\overline{h_{k_1,m_1} h^*_{k_2,m_2}} = r(\Delta m) = J_0\left(2\pi f_D T \Delta m\right), \quad (6)$$

where

$$\Delta m = m_1 - m_2 + (k_1 - k_2)(N_Q + 1). \quad (7)$$

Here $J_0(\cdot)$ is Bessel function of the first kind of order zero, and f_D is Doppler spread. Overline denotes ensemble averaging, and * denotes complex conjugation.

Since the average power of all symbols is equal to one, and channel gains have unit variance, signal-to-noise ratio (SNR) per bit is

$$\frac{E_b}{N_0} = \frac{1}{\mathrm{SE} \cdot P_n}. \quad (8)$$

3. Reception Methods

3.1. Methods for AWGN

In the case of slow fading, reception method presented in [6] can be used. In this method, DPSK is demodulated, and obtained hard decisions are used to estimate a constant channel gain for QAM symbols. Mathematical description of the method has the following form:

1) DPSK demodulation:

$$\hat{d}_k = \left[\frac{M_D}{2\pi} \arg\left(x_{k,(N_Q+1)} x^*_{k,0} \right) \right]. \quad (9)$$

Here [ ] denotes rounding to the nearest integer.

2) Hard decision $\hat{d}_k$ gives a pair of pilot symbols:

$$p_0 = 1, \quad p_1 = \exp\left(j \frac{2\pi}{M_D} \hat{d}_k \right). \quad (10)$$

3) To estimate CSI for QAM symbols, several approaches can be used. In all considered cases the estimate is obtained as a linear combination of the following form:

$$\hat{h}_{k,m} = x_{k,0}\, p_0^*\, w_{0,m} + x_{(k+1),0}\, p_1^*\, w_{1,m}. \quad (11)$$

The weighting factors $w_{0,m}$, $w_{1,m}$ depend on particular estimation method. The following methods are considered:

a) Maximum likelihood (ML) estimation of a constant channel gain (this formula was used in [6] for AWGN channel):

$$w_{0,m} = w_{1,m} = \frac{1}{2} \quad \forall m = 1, \ldots, N_Q. \quad (12)$$

b) Linear interpolation:

$$w_{0,m} = 1 - \frac{m}{N_Q + 1}, \quad (13)$$

$$w_{1,m} = \frac{m}{N_Q + 1}. \quad (14)$$

c) Optimal Wiener filter (this method requires the knowledge of the channel gain correlation properties (6), (7)):

$$w_{0,m} = \frac{r(m) - r(N_Q + 1 - m) r(N_Q + 1)}{1 - (r(N_Q + 1))^2}, \quad (15)$$

$$w_{1,m} = \frac{r(N_Q + 1 - m) - r(m) r(N_Q + 1)}{1 - (r(N_Q + 1))^2}. \quad (16)$$

Here, high SNR is assumed, so the influence of noise is ignored.

4) QAM symbols are demodulated with account for CSI obtained at the previous step:

$$\hat{a}_{k,m} = \arg\min_{C \in \mathcal{C}} \left(|x_{k,m} - \hat{h}_{k,m} C| \right). \quad (17)$$

It should be noted that DPSK demodulation in the case of fast fading produces an error floor [8]. These errors that do not cease even at high SNR would lead to large errors in CSI estimation for QAM symbols and therefore an error floor would also appear for QAM demodulation.

3.2. GLRT Reception

According to the generalized likelihood ratio test (GLRT) detector description in [4], it computes the joint ML estimate of the channel and the transmitted signal. In the case of AWGN channel, the decision formula is simple but it may require an exhaustive search over signal set. In the case of fast fading, where estimated channel gain is not constant, GLRT reception procedure becomes highly complicated. To implement it, for every possible symbol sequence of Figure 3 the following calculations should be done:

1) CSI estimation. As it is based on a complete hypothesized symbol sequence, it should be essentially some sort of smoothing. To get an optimal estimate, channel model is needed.
2) Squared Euclidean distance calculation between received sequence of samples and hypothesized symbol sequence multiplied by CSI estimates.

After calculating the distances for all $M = M_D M_Q^{N_Q}$ possible symbol sequences, the sequence that gives the minimum distance is selected as a final decision.

The necessity of exhaustive search through the complete signal set makes this method unfeasible for almost all combinations of M_D, M_Q, and N_Q. Therefore, some simplifications are required to make this approach suitable for practical use. One such modification is presented below.

3.3. Proposed Method

The main idea of the proposed reception method is that the search and CSI estimation are performed only over possible DPSK symbols, while for QAM symbols hard decisions are used. Such approach is computationally feasible and at the same time it involves all symbols of the block into the process of selecting DPSK hypothesis. Therefore, it can be treated as a simplified approximation of GLRT. Computational cost of this method is $M_Q^{N_Q}$ times lower than that of GLRT.

Mathematical description of the method has the following form. We consider all M_D DPSK hypotheses sequentially, and for every integer number $d = 0, 1, \ldots, M_D - 1$ perform the following steps:

1) From the hypothesis d, a pair of pilot symbols is formed similar to (10):

$$p_0 = 1, \quad p_1 = \exp\left(j \frac{2\pi}{M_D} d \right). \quad (18)$$

2) These pilot symbols are used to estimate the CSI for both DPSK and QAM symbols. As the channel gain between DPSK samples is not constant, various interpolation techniques can be used here, such as (12), (13) and (14), or (15) and (16). In the simulations, Wiener filtering (15), (16) was used. As a result of this step, we obtain CSI estimations $\hat{h}_{k,m}$, $m = 0, 1, \ldots, N_Q + 1$.

3) QAM symbols are demodulated with account for obtained CSI according to (17).
4) Squared Euclidean distance between received samples and hard decisions for both DPSK and QAM symbols is calculated:

$$u_d^2 = \sum_{m=1}^{N_Q} |x_{k,m} - \hat{h}_{k,m}\hat{a}_{k,m}|^2 + |x_{k,0} - \hat{h}_{k,0}|^2 + |x_{k,(N_Q+1)} - \hat{h}_{k,(N_Q+1)}p_1|^2. \quad (19)$$

In this formula, two last terms correspond to the DPSK samples.

After calculating u_d^2 for all M_D DPSK hypotheses, the result that provides the minimum u_d^2 is selected:

$$\hat{d}_k = \arg\min_d u_d^2, \quad (20)$$

and demodulation results corresponding to the selected hypothesis are stored.

To illustrate this idea, an example will be considered with the following signal parameters: $M_D = 4$, $M_Q = 16$, $N_Q = 16$, $E_b/N_0 = 20$ dB, $f_D T = 10^{-2}$.

All $N_Q + 2 = 18$ samples of one signal block are shown in Figure 4. Two DPSK samples are marked by the green rectangles. It is seen that the angle between these samples is close to $135°$, hence two DPSK hypotheses ($90°$ and $180°$ phase shift) are almost equiprobable.

Figure 5 illustrates the steps presented above, for the first DPSK hypothesis $d = 0$. Red dots show N_Q 16QAM constellations rotated and scaled by the CSI estimations $\hat{h}_{k,m}$. Red circles mark the selected QAM points $\hat{a}_{k,m}$, and blue lines show the distances $|x_{k,m} - \hat{h}_{k,m}\hat{a}_{k,m}|$.

Figures 6–8 are similar to Fig. 5, they correspond to the three remaining DPSK hypotheses ($d = 1, 2, 3$).

Calculations of u_d^2 in this example produce the results presented in Table 1. It is seen that the best result is obtained when $d = 2$ (DPSK rotation by $180°$). It should be noted that the ordinary DPSK demodulation in this example would give $\hat{d}_k = 3$ (DPSK rotation by $-90°$). This fact shows that the proposed reception method allows to exploit the information about the QAM signal constellation without the exhaustive search over all possible symbol combinations.

4. Simulation Results

The first simulation was performed to compare different reception methods. The resultant dependencies of bit error rate (BER) on the average SNR per bit are shown in Figure 9 for the following signal parameters: $M_D = 8$, $M_Q = 16$, $N_Q = 4$. Two values of Doppler spread were simulated: $f_D T = 10^{-2}$ and $5 \cdot 10^{-3}$. Plot legend indicates Doppler spread values and the reception methods:

- **Wiener:** reception algorithm for AWGN, optimal filtering with the weights defined by (15) and (16);
- **DPSK search:** proposed method (18)–(20).

Reception algorithm for AWGN was also simulated with linear interpolation (13), (14) and estimation of a constant gain (12), but the influence of CSI estimation method on the BER results proved to be negligible. Therefore, the results are shown only for Wiener filtering.

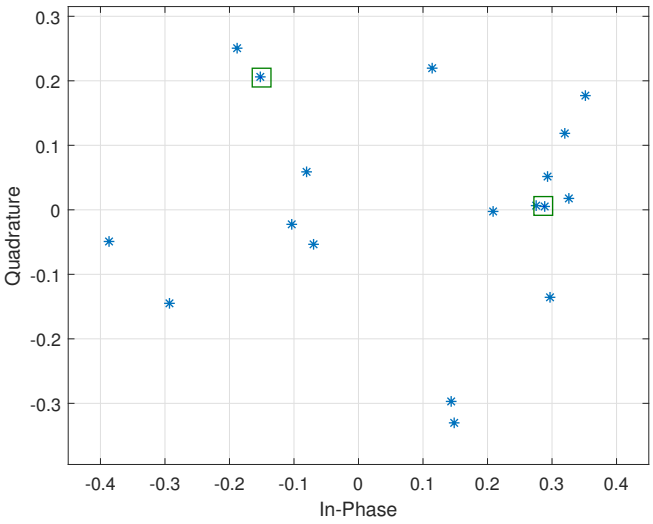


Figure 4. Example: received samples of one signal block

TABLE 1. VALUES OF u_d FOR PRESENTED EXAMPLE

d	0	1	2	3
u_d^2	0.2724	0.3405	**0.05491**	0.09636

The curves show that, as expected, reception method intended for AWGN channel produces an error floor that increases with Doppler spread. The proposed reception algorithm, due to inclusion of QAM symbols into the process of DPSK demodulation, makes the error floor several orders of magnitude lower though does not remove it completely.

The purpose of the second simulation was to compare the influence of Doppler spread on BER for signals with different combinations of parameters that provide the same SE. The proposed reception method was used.

Figure 10 shows the dependence of BER on Doppler spread at $E_b/N_0 = 20$ dB for several combinations of M_D and M_Q with $N_Q = 1$ and SE $= 4$ bits per symbol.

The curves are labeled as "DxQy", where $x = M_D$ and $y = M_Q$. When $M_D = 1$, it means the absence of DPSK modulation, i. e., the use of the ordinary pilot symbols.

It is seen that lower M_D generally improves tolerance to high f_D values, but for slowly changing channels (lower f_D values) DPSK-modulated pilots can provide lower BER.

For every f_D value, an optimal M_D can be found. Below, appropriate approximate ranges of $f_D T$ are listed for various M_D:

- $M_D = 1$ (unmodulated pilots): $f_D T > 4 \cdot 10^{-2}$;
- $M_D = 2$: $f_D T = (2 \dots 4) \cdot 10^{-2}$;
- $M_D = 4$: $f_D T = (1 \dots 2) \cdot 10^{-2}$;
- $M_D = 8$: $f_D T < 10^{-2}$;
- $M_D = 16$: not optimal $\forall f_D T$.

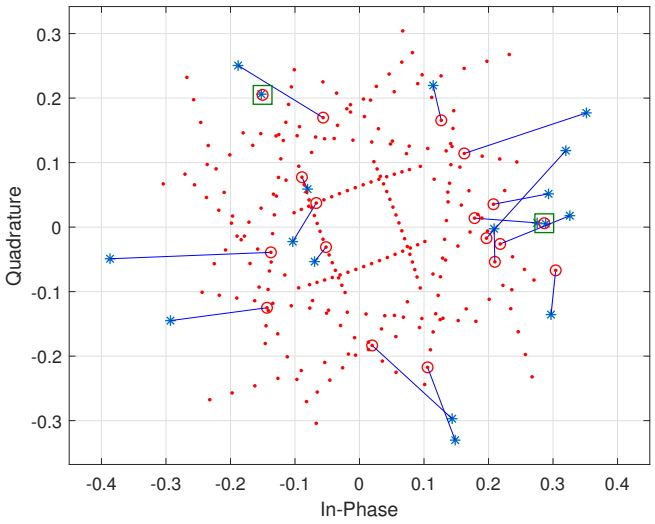


Figure 5. Results of CSI estimation and QAM demodulation for the DPSK hypothesis $d = 0$

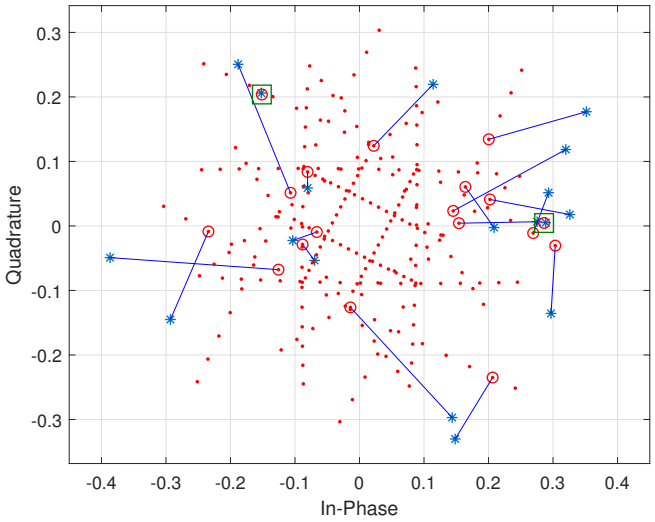


Figure 6. Results of CSI estimation and QAM demodulation for the DPSK hypothesis $d = 1$

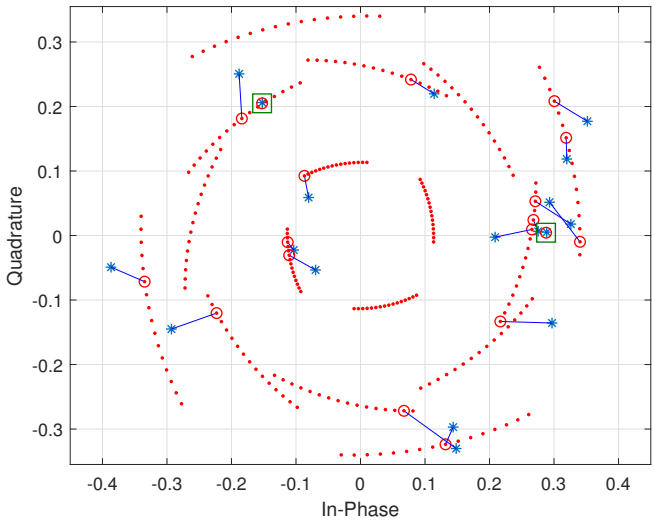


Figure 7. Results of CSI estimation and QAM demodulation for the DPSK hypothesis $d = 2$

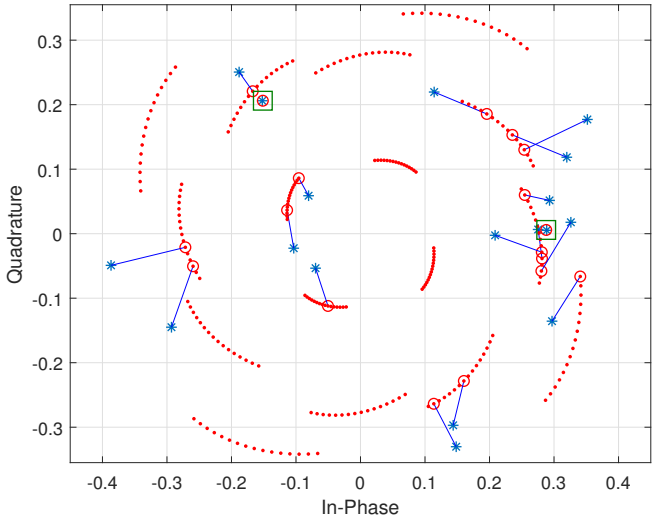


Figure 8. Results of CSI estimation and QAM demodulation for the DPSK hypothesis $d = 3$

Figure 10 also shows that for low values of f_D, combination $M_D = 8$, $M_Q = 32$ is optimal. It coincides with optimization results presented in [6] for AWGN channel.

In the Figures 11 and 12, similar results are shown for $N_Q = 2$ and $N_Q = 4$, respectively. In these cases, some values of M_D do not allow to obtain SE $= 4$, so the most close values were used. The curves are labeled as "Dx(Qy)^{N_Q}", where $x = M_D$ and $y = M_Q$.

The general behavior of the curves is the same as in Figure 10 (lower M_D allows to tolerate higher f_D), but potential BER decrease due to the use of higher M_D drops with N_Q: BER reduction factor is about 5 for $N_Q = 1$, about 2 for $N_Q = 2$ and about 1.64 for $N_Q = 4$. Consequently, the maximum gain from the use of the proposed scheme can be achieved for small values of N_Q.

5. Conclusion

The proposed reception method has allowable computational cost and significantly (several orders of magnitude) reduces error floor in channels with fast fading. With this reception method, it is possible to improve power efficiency of the transmission by using DPSK-modulated pilot symbols. Optimal DPSK constellation size decreases with Doppler spread. The achievable BER decrease due to the use of DPSK-modulated pilot symbols drops with the number of inserted QAM symbols N_Q. In the case of $N_Q = 1$, BER can be reduced by the factor of 5.

Possible directions of future research:

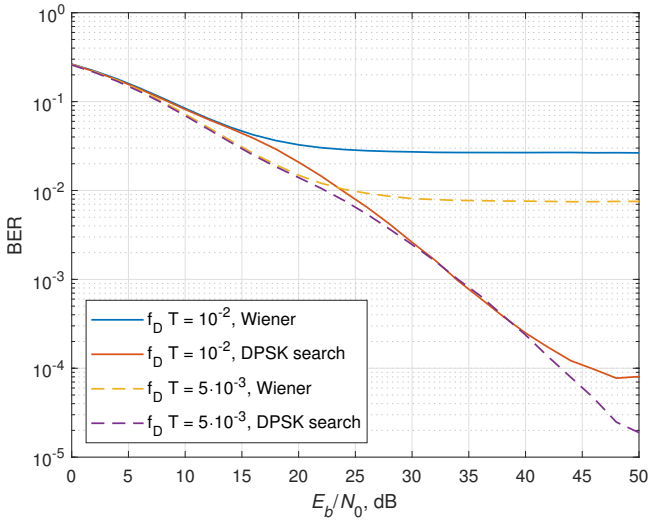


Figure 9. BER vs. average SNR per bit, $M_D = 8$, $M_Q = 16$, $N_Q = 4$

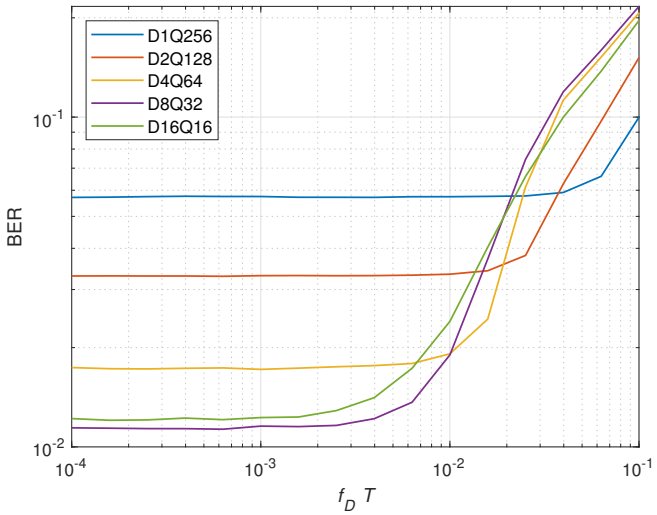


Figure 10. BER vs. Doppler spread, $N_Q = 1$, SE $= 4$, $E_b/N_0 = 20$ dB

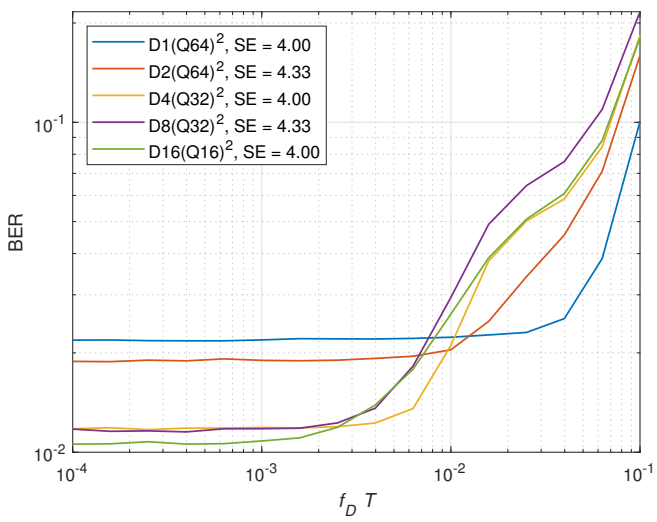


Figure 11. BER vs. Doppler spread, $N_Q = 2$, SE ≈ 4, $E_b/N_0 = 20$ dB

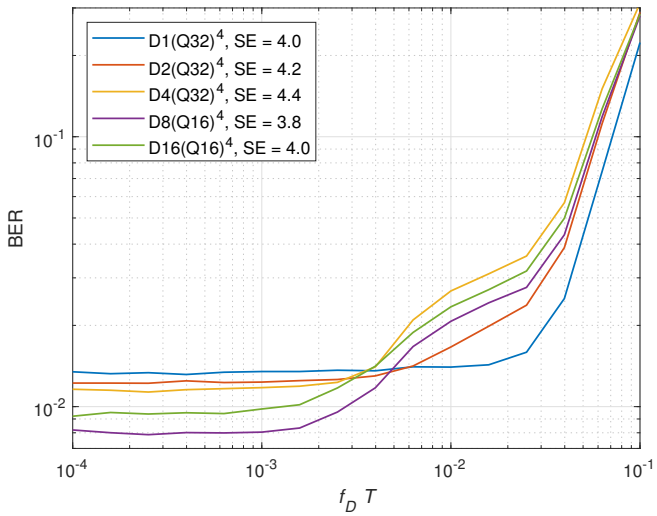


Figure 12. BER vs. Doppler spread, $N_Q = 4$, SE ≈ 4, $E_b/N_0 = 20$ dB

- analytical estimation of the BER performance of the proposed method;
- modification of proposed reception method to calculate soft decisions for this modulation scheme;
- extension of this approach to multicarrier and spatially multiplexed systems.

References

[1] A. B. Sergienko, "Noncoherent reception of short PSK data packets with pilot symbols," in *European Wireless 2016; 22th European Wireless Conference*, May 2016, pp. 1–6.

[2] ——, "Semi-blind approach to reception of short QAM data packets with pilot symbols," in *2016 XV International Symposium Problems of Redundancy in Information and Control Systems (REDUNDANCY)*, Sept 2016. doi: 10.1109/RED.2016.7779348 pp. 137–141.

[3] J. Proakis, *Digital Communications*, ser. Electrical engineering series. McGraw-Hill, 2001. ISBN 978-0-07-232111-1

[4] D. Warrier and U. Madhow, "Spectrally efficient noncoherent communication," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 651–668, Mar 2002. doi: 10.1109/18.985996

[5] A. Saci, A. Al-Dweik, A. Shami, and Y. Iraqi, "One-shot blind channel estimation for OFDM systems over frequency-selective fading channels," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5445–5458, 2017. doi: 10.1109/TCOMM.2017.2740925

[6] A. B. Sergienko, "DPSK-QAM combination as a signal set for spectrally efficient noncoherent communication," in *2018 IEEE East-West Design Test Symposium (EWDTS)*, 2018. doi: 10.1109/EWDTS.2018.8524652 pp. 1–6.

[7] R. H. Clarke, "A statistical theory of mobile-radio reception," *The Bell System Technical Journal*, vol. 47, no. 6, pp. 957–1000, 1968. doi: 10.1002/j.1538-7305.1968.tb00069.x

[8] M. Simon and M. Alouini, *Digital communication over fading channels: a unified approach to performance analysis*, ser. Wiley series in telecommunications and signal processing. John Wiley & Sons, 2000. ISBN 978-0-471-31779-1

An IoT based Real-time Data-centric Monitoring System for Vaccine Cold Chain

Raisa Tahseen Hasanat
Department of Electronics and Telecommunication Engineering
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
tahseen.hasanat.ete@ulab.edu.bd

MD. Arifur Rahman
Team AOS
Dhaka, Bangladesh
ar13101085@

Nafees Mansoor
Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
nafees.mansoor@ulab.edu.bd

Nabeel Mohammed
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
nabeel.mohammed@northsouth.edu

Mohammad Shahriar Rahman
Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
shahriar.rahman@ulab.edu.bd

Mirza Rasheduzzaman
Department of Electronics and Telecommunication Engineering
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
mirza.rasheduzzaman@ulab.edu.bd

***Abstract*—Vaccination is the assured way of gaining immunization against many life-threatening diseases. However, the vaccine outreaches in developing and undeveloped countries are very limited due to lack of proper management of the cold chain system. This paper presents a real-time data–centric cold chain monitoring system for the continuous monitoring of the vaccine distribution and transportation process. The proposed system provides the unique feature of creating and managing individual trips for vaccine transportation process along with the regular supervision of temperature and humidity of the carrier. Moreover, the hardware and software components for the system also track the location of the carrier. This proposed system can be particularly highly effective in increasing vaccine coverage in the remote regions. This is because the proposed system enables the remote monitoring of the entire process and ensure transparency in the distribution process.**

Keywords— IoT, Monitoring System, Vaccine Cold Chain.

I. INTRODUCTION

Immunization by vaccines is widely acknowledged for controlling and eliminating a large number of infectious diseases and is also one of the most cost-effective public health interventions. According to UNICEF, vaccines are saving 2-3 million lives every year. However, vaccine outreach is still very limited in the developing and undeveloped countries. In 2018 alone, 13.5 million children did not receive routine immunization and 1.5 million lives are lost every year from diseases that can be prevented by vaccines [1].

Even though different factors are responsible for this low outreach of vaccines, breach of the vaccine cold chain is the biggest contributing factor [2]. Vaccines are extremely sensitive to temperatures. The World Health Organization has fixed the temperature range for vaccine storage and transportation as 2-8°C and vaccines completely lose potency if they are exposed to temperatures beyond this range even for short durations. This is why maintaining the cold chain system from the point of manufacture till the point of administration is very important. However, various physical, geographical and socio-economic factors in the developing countries hamper the smooth management of the vaccine cold-chain system resulting in the loss of almost 50% vaccines annually [3]. WHO has a number of standardized devices and guidelines for monitoring the cold chain; but in the undeveloped countries, about 31% of these devices were non-functional and several of the units were too old for use [4]. Most of the people in the undeveloped and developing countries are not sufficiently trained for using these devices and often times there is no transparency and no routine monitoring in the cold chain system. This leads to the wastage of almost 39.54% vaccines at the session sites. Moreover, while transporting, the vaccines are mostly carried in cold boxes using ice packs and cold water packs. The vaccines often freeze below the necessary temperature range rendering them useless and even harmful. Again, because of the lack of accountability, vaccines even get lost or stolen during the journey to the health centers. All these contribute to the loss of almost 30% vaccines during transportation [5].

Considering these existing problems in maintaining the cold chain, this paper introduces a real-time, data-centric vaccine cold chain monitoring system, which monitors the temperature and location of the vaccine carriers and sends necessary notifications and text messages to the healthcare supervisors accordingly. The corresponding hardware designs for the system have been previously developed which consists of a thermoelectric based vaccine carrier, a monitoring module containing the necessary temperature and humidity sensors and a communication unit which is responsible for providing the location information as well as for sending text notifications and sharing data on the web server [6]. This system ensures that transparency is maintained and that the vaccines are monitored routinely throughout the transfer process. Unlike traditional

Funding from ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of Bangladesh is gratefully acknowledged.

systems, the entire monitoring process is carried out automatically, hence there is no space for human errors or negligence. The different design aspects and features of the system are described in details in this paper.

This paper is organized as follows. The background study and the limitations and strengths of the existing systems are mentioned in section II. The different features of the proposed system is presented in section III. Section IV contains the implementation framework of the system and the conclusion and future scopes are highlighted in section V.

II. Background Study

Over the years, numerous researches have been carried out to develop web and mobile based applications which can assist in increasing vaccine coverage and monitoring the cold chain. Among the various works done, most of the mobile applications that have been developed are largely responsible to collect data about vaccine coverage in the hard to reach areas of developing countries and are concerned with monitoring the routine administration of vaccines [7, 8, 9]. However, as previous discussions indicate, the proper coverage of vaccines in these remote regions can only be ensured when the cold chain of the vaccine supply system performs properly. So even though these applications can successfully determine the rate of vaccine outreach, they do not work to increase this coverage rate.

Moreover, there exists a few web based systems as [10, 11] which allows to monitor the cold chain performance by reporting the refrigeration status of the various vaccine refrigerators, cold rooms and boxes. However, these systems are only suitable for use in vaccine cold rooms and vaccine refrigerators at the session sites. These are not adapted for use during the transportation period of the vaccines where the cold chain breach is most common. In [12] a cold chain monitoring system is devised based on FonAstra which a sensing system that uses cellular network to send temperature and location data via text messages. This data is then stored in a database which is accessible through a web browser. A similar work has also been done in [13] where the system sends the data through WI-FI transmitters to the monitoring authorities who can check this data through Sensor cloud. It does not send any separate SMS notification. Both of these systems are only online based applications and these also have a basic limitation that these do not include the function of allocating separate IDs to the large number of vaccine carriers that are often sent out together on different trips and also cannot identify from which carrier or which vaccine delivery trip the definite data is being sent from. So these in practice can only be used for individual carriers or single trips at a time. The system discussed in [14] is a slightly improved version of the two previous systems. This monitoring system, sends the temperature and location data to the supervisor as SMS and also sends to a cloud based web service. The cloud application receives acknowledgements from the supervisor's mobile phone and marks the vaccine carriers as safe or unsafe accordingly. Unlike the previous systems, this monitoring system can identify individual boxes by separate IDs but do not have the ability to differentiate between the various delivery trips being made at the same time.

This system thus have been designed considering the limitations of the existing works. Unlike the previous works, the monitoring system is based on both a web server and a mobile application. The mobile application has a user friendly interface which allows easy access to the health care workers. Moreover, this work has a number of distinctive features which were absent in the previous works; most important of these being: assignment of individual ID numbers to all the vaccine carriers and creating and tracking every individual trip that is made to the remote health centers.

III. Proposed Model

A general idea about how the monitoring system functions is depicted in the block diagram in Fig. 1.

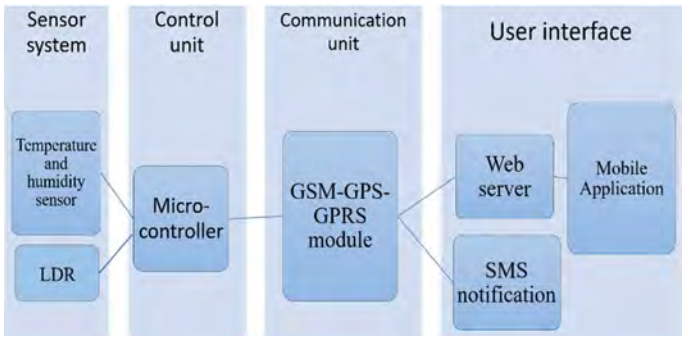


Fig. 1. Monitoring system block diagram

A. Trip details and vaccine carrier information

The first feature that sets this system apart from the previous works is that at the beginning of any vaccine transportation trip, a new trip is created through the web server of the system which can be accessed by the healthcare supervisor. This contains all the necessary details about an individual trip, including the current status of a trip, the trip route, the details about the type and amount of vaccines being carried and details about the healthcare workers on the trip. The entire trip can also be tracked by the healthcare supervisors or admin while it is active. While tracking an individual trip, the admins can access information about the current location of the carrier, the current temperature and humidity readings of the vaccine carrier chamber as well as the general information about the healthcare worker and the vaccines being carried. However, while tracking there is no provision for selecting the best route for the vaccine transfer, rather, the carrier will follow the route pre-defines before the journey. Algorithm 1 describes the process of searching and tracking an active trip.

Algorithm 1: Algorithm for searching and tracking a trip

```
1:  start
2:  tripData = getTripInformationFromRoute();
3:  trip = getTripDetailsById(tripData.id);
4:  if trip == null then
5:      trip = makeTrip(tripData);
6:      trip.locationList=tripData.locationList
7:      trip.userList=tripData.userList
8:  end
9:  allLocation = trip.getAllLocation();
10: allUser =trip.getAllUserList();
```

```
11:  isAllLocationVisited = true;
12:  for every location in a allLocation
13:      if notlocation.isVisited() then
14:        isAllLocationVisited = false;
15:      end
16:  endfor
17:  if isAllLocationVisited then
18:     trip.isCompleteTrip = true;
19:  else
20:    trip.isCompleteTrip = false;
21:  end
22:  end
```

Also, each of the vaccine carrier in the trips are assigned a unique ID which allows them to keep track of all the carriers individually as well.

B. Continuous monitoring

The vaccine carriers used in the system are equipped with a temperature and humidity sensor that continuously monitors the temperature and humidity of the carrier chamber and the monitored data is sent to a microcontroller unit which records the temperatures and sends the data at a regular interval to the communication unit. Moreover, an LDR is placed in the carrier that helps to determine if the vaccine carrier has been opened anytime during the journey. If the carrier is opened, the LDR value crosses a threshold level which causes the communication unit to send urgent notifications regarding opening of the carrier. This ensures that no vaccine can be stolen or removed from the carrier during the transportation process.

C. Location Tracking

In this system, the communication unit consisting of the GSM-GPS-GPRS module is responsible for providing the location information of the vaccine carrier along with determining the date and time of the trip being made. The module determines the GSM location of the carrier and this enables the supervisor to locate the position of any vaccine carrier at any given time to ensure that the pre-defined route for vaccine transportation is being followed.

D. Regular and Urgent Notifications

In the proposed system, at definite regular intervals, notifications about each individual box are sent to the supervisor as well as to the health care individual assigned to the particular trip. These notifications contains the carrier temperature, humidity and location information along with the time stamp. Apart from the regular notifications, some urgent notifications are also sent out whenever the carrier temperatures fall beyond the range of 2-8°C or if the carrier has been opened during the trip. Both the regular and urgent notifications are sent to the users via SMS as well as through the mobile application, because these carriers are designed to be used even at the most remote areas. Due to the lack of proper infrastructure, it is possible that at any point of the transportation process, internet connections might not be available or might be weak which can cause a delay in receiving notifications from the mobile application. However, basic mobile networks are mostly available in all areas and SMS notifications are more likely to be received.

IV. Framework Implementation

A. Architectural Design

The architectural design of the proposed system is based on a three-tier approach where the first tier is the client tier or presentation tier, the second tier is the middle tier followed by the database tier. This three tier architecture allows to integrating the different users and their functionalities on one platform. This enables the proposed system to provide more flexibility to its applications.

1) The Client Tier: The first tier in the architecture is the client tier which allows the users to interact with the platform. In this case, the web browser and the smart phone's touch screen and display interface act as the client tier which is utilized by all users connected to the system. The graphical user interfaces of the application operates in this tier and it allows the users to perform the different functions like creating an account, logging in, tracking a current trip and so on. This proposed system provides a very simple and user friendly GUI, which will allow the healthcare workers with minimum training and education to easily operate the application. Fig. 2 and Fig. 3 presents the mobile and web based interfaces of the proposed monitoring system respectively.

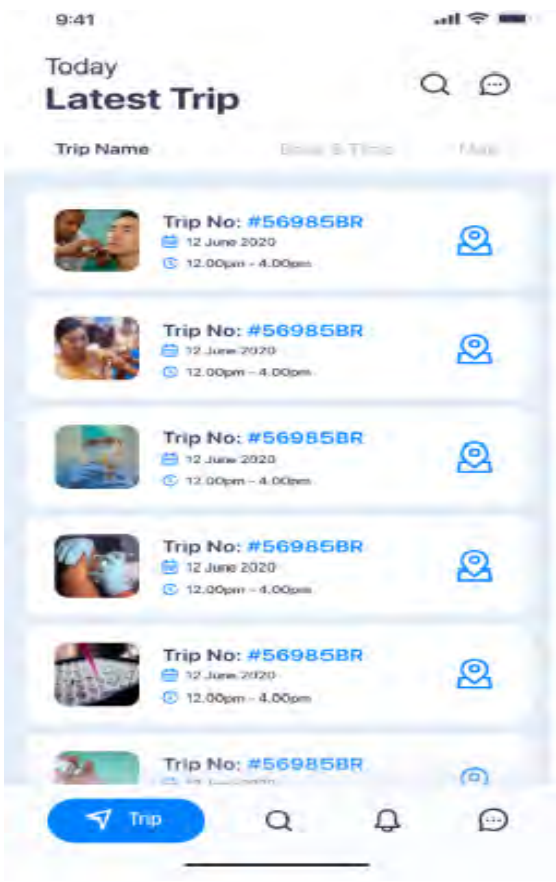


Fig. 2. Mobile user interface

Fig. 3. Monitoring system homepage

2) The Middle Tier: In this system, the middle tier consists of the core logic of the application. As mentioned before, the client tier assembles and diplays data from the user, while the database performs the storage and retrieval of data. Most of the remaining functions are performed by the middle tier where it determines the content and structure of the data to be displayed to the users and processes the user input. This middle tier acts a merger between the other two tiers. The inputs generated by the users are formed into queries and interpreted by the data base as read or write functions. In this proposed system, the web server is a core component of the middle tier.

3) The Database Tier: The final tier of the architectural design is the database tier which is responsible for the storage and retrieval of all data from the system. The database of the proposed system is designed based on the entity relationship diagram during the system design phase and it consists of seven database tables.

B. Role of the Users

The proposed system comprises of two types of users: the healthcare supervisors who are responsible for creating the trips and monitoring the entire transportation process of the vaccines and the healthcare workers who are assigned for individual trips to transport the vaccines to the centers. All the users are required to be registered in the platform to use it. Next based on the type of the user, they are directed to the respective dashboards where they can perform their individual functions as depicted in Fig. 4.

1) Healthcare Supervisors/ Admin: The first role of the supervisors or admins is to create an account and log into the account using their emails. Supervisors are allowed to login both to the web and mobile applications. Next the user can access a number of features on the system as depicted in Fig. 4. This includes creating, updating, viewing or deleting trips and vaccine information and tracking the locations and access details about all ongoing trips. During an ongoing trip, the supervisors also receive routine notifications through the mobile app about the temperature conditions of the carrier and whether the carrier has been opened or not and can take actions accordingly.

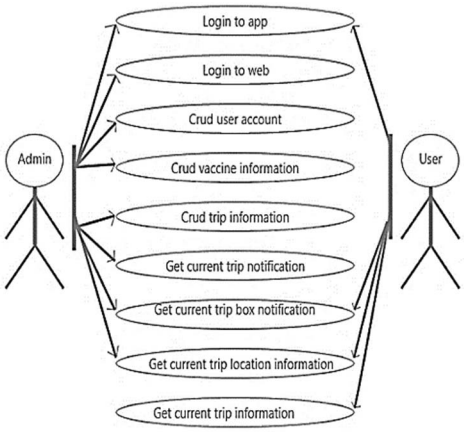


Fig. 4. Use case diagram

2) Healthcare Workers: Similar to the supervisors, the first role of the workers assigned to the trips is to create and log in to their respective accounts. The workers have access only to the mobile applications and they can view the details about their current trips only. They receive regular notifications about the temperature conditions and can view the location and route information of their current trips.

3) Physical Design: After completing the system analysis and understanding the role of each user, the logical designs of the systems are created. Each process in the system was modeled using Data Flow Diagrams (DFD) and to get better understanding of the flow of data through the system Entity Relationship Diagrams (ERD) were constructed. Moreover, sequence diagrams were utilized to determine the roles and functions of each user. The main concerns of this design level was thus to determine the work flow of the entire system and how the outputs will be presented.

C. Data Flow of the Application

1) Sign in and Sign up: Irrespective of the type, all the users must go through the process of signing up and logging in. The process of creating an account and logging in is very simple. For creating an account, the user is required to click on the Register button. This will prompt the user to create an account by providing the necessary information and click submit. This sends the details of the user to the database. Once the sign up process is successful, the users can login to their accounts by entering the necessary credentials. After logging in the different features can be accessed based on the type of the user.

2) 4.3.2 Creating and Searching for a Trip: One of the features of the system is the ability to create individual trips before a specific vaccination transfer process is started. For this purpose the user who is admin or supervisor is required to access the feature by clicking on the “create new trip” button on the dashboard. This provides a form requesting different information regarding the trip and the vaccines being carried.

Similar to creating a trip, only an admin can search for both current and previously completed trips by using the search option on the dashboard. The trip name entered by the user is matched against the database to retrieve and present the details of that particular trip. The dashboard of the admin also by default contains a list of ongoing trips and the details of a trip can be viewed simply by clicking on the list item. Similar processes are also to be followed by the healthcare workers, the only difference being that they are allowed to view only the trip that they are currently assigned in.

3) Adding New Vaccine types: For adding details about new vaccines to the database, a similar process is followed as creating a trip. The admin accesses a form by clicking on the "create new vaccine" button and by providing the necessary information in a form, the details of the vaccine are saved in the database. The unique vaccine ID assigned to each vaccine created in this process is utilized later in creating a trip.

V. Conclusion and Future Works

An app based monitoring system for vaccine cold chain is proposed in this system. The proposed system ensures that the vaccine cold chain is monitored continuously during the entire transportation process to health centers and as the system is completely automated, it ensures transparency and efficiency in the whole process. The system can be extremely beneficial when utilized to track and monitor the vaccine transfer processes in the remote areas in developing and undeveloped countries where the vaccine outreach is minimum due to lack of proper management of the cold chain systems. Even though there are no visible drawbacks of the system, there are still some future scopes in improving the overall efficiency of the application by making it as user friendly as possible to be easily usable by the untrained field workers and add features as necessary.

Acknowledgment

The authors would like to gratefully acknowledge the funding from the ICT Innovation Fund under ICT Division, Ministry of Posts, Telecommunications and Information, Government of Bangladesh.

References

[1] Unicef and World Health Organization, "Progress and Challenges with Achieving Universal Immunization Coverage : Estimates of Immunization Coverage," no. July, pp. 1–18, 2019.

[2] A. Ashok, M. Brison, and Y. LeTallec, "Improving cold chain systems: Challenges and solutions," *Vaccine*, vol. 35, no. 17, pp. 2217–2223, 2017.

[3] World Health Organisation, "Monitoring vaccine wastage at country level," *World Heal. Organ.*, pp. 21–27.

[4] Federal Ministry of Health Ethiopia(FMoH), "Ethiopia National Expanded Program on Immunization, Comprehensive Multi - Year Plan 2016 – 2020.," pp. 1–115, 2015.

[5] A. PEEPLIWAL, "Cold Chain: A Lynchpin of National Immunization Program," *J. Pharm. Technol. Res. Manag.*, vol. 5, no. 1, pp. 77–104, 2017.

[6] R.T. Hasanat, N. Mansoor, M.S. Rahman, M. Rasheduzzaman, 2020. Development of Monitoring System and Power Management for an IoT based Vaccine Carrier. In *ICED 2020.*

[7] A. Katib, D. Rao, P. Rao, K. Williams, and J. Grant, "A prototype of a novel cell phone application for tracking the vaccination coverage of children in rural communities," *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 215–228, 2015.

[8] M.J. Uddin, *et al.*, "Use of mobile phones for improving vaccination coverage among children living in rural hard-to-reach areas and urban streets of Bangladesh," *Vaccine*, vol. 34, no. 2, pp. 276–283, 2016.

[9] D. Hansen, Niels, et al., "Time-Series Adaptive Estimation of Vaccination Uptake Using Web Search Queries." Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017.

[10] R. Anderson, *et al.*:"Supporting immunization programs with improved vaccine cold chain information systems," *Proc. 4th IEEE Glob. Humanit. Technol. Conf. GHTC 2014*, pp. 215–222, 2014.

[11] R. Anderson, J. Lloyd, and S. Newland, "Software for national level vaccine cold chain management," *ACM Int. Conf. Proceeding Ser.*, pp. 190–199, 2012.

[12] R. Chaudhri, G. Borriello, and R. Anderson, "Pervasive Computing Technologies to Monitor Vaccine Cold Chains in Developing Countries," *IEEE Pervasive Comput.*, vol. 11, no. 3, pp. 26–33, 2012.

[13] A. Mohsin, and S.S. Yellampalli, "IoT based cold chain logistics monitoring," *IEEE Int. Conf. Power, Control. Signals Instrum. Eng. ICPCSI 2017*, pp. 1971–1974, 2018.

[14] A. Parthasarathy, J. Chinnappa and P.S. Ramkumar, "A Cloud-based Vaccine Cold-Chain Monitoring System," *ASCI Journal of Management*, vol. 46, pp. 129-136, 2017.

RFID-Based Navigation of Subway Trains

Alexander M. Kostrominov
D. Sc., Professor at Department of Electrical Communication, Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
triak@grozon.spb.ru

Oleg N. Tyulyandin
Ph. D. student at Department of Electrical Communication, Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
tyulyandin@vklay.ru

Alexander B. Nikitin
D. Sc., Professor at Department of Automation and Remote Control on Railways, Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
nikitin@crtc.spb.ru

Michael N. Vasilenko
D. Sc., Professor at Department of Automation and Remote Control on Railways, Emperor Alexander I St. Petersburg State Transport University,
St. Petersburg, Russia
vasilenko.m.n@gmail.com

Alexander T. Osminin
D. Sc., Professor, Vice-chairman Joint Scientific Council of JSC Russian Railways
Moscow, Russia
at@osminin.com

***Abstract*—Most tasks related to automatic control of subway trains require availability of accurate train coordinates. Odometry is the simplest method of positioning, but it tends to accumulate errors. We suggest using radio frequency (UHF-RFID) positioning in addition to odometry, which will allow to nullify the errors and calibrate the odometer sensors for more accurate positioning of subway trains. Active readers and antennas shall be installed on head cars, and passive RFID tags storing their coordinates shall be put on tunnel walls. This paper addresses the topical issue of reliability of RFID tag scanning by equipment of various generations depending on train speed, radio visibility zone and programmed time of tags scanning. A method to evaluate the accuracy of precision RFID-based positioning of subway trains is proposed. Basing on train speed and parameters of reader to-tag communication, it becomes possible to give sufficiently accurate evaluation of a reader antenna position relatively to the RFID tag location at the first scanning. The obtained results are universally valid and can be used both on surface rail transport, and in other related fields.**

Keywords—RFID, navigation, positioning, accuracy, reliability, subway trains.

I. Introduction

The accurate positioning (navigation) of subway trains becomes more critical when implementing automatic train control. Accurate positioning is required not only for target braking at stations with platform screen doors, but also for continuous determining current coordinates of a train as the base of proper performance of the entire automatic train control system [1, 2].

The main navigation tool in subways is odometry (it uses wheel sensors to show travel distance). Current train position on a line is tracked against a certain point (dead reckoning), for instance, departure from a stub track or the first station [3]. However, such method of positioning has multiple disadvantages: it is highly dependent on accuracy of sensor calibration, difference of the distance traveled by a wheel pair in curves and spinning of wheels while accelerating and slowing down. There exist mathematical methods to improve the accuracy of positioning which involve the use of wheel sensors [4, 5], but they do not allow to eliminate the errors completely. Refined odometry readings require additional tools.

Positioning of subway trains can use Radio Frequency Identification. Such technology is implemented in St. Petersburg Metro, where some lines have been equipped with automatic train control system that uses the UHF RFID technology [6–10]. In such is case, automatic train control commands are recorded in memory of the floor-level RFID tags put on tunnel walls. Coordinates of the tunnel tags are stored in their memory, too. Communication between trains and tags goes through a radio channel using readers installed in head cars of the trains.

The navigational constituent of such system is particularly important. Such system can be used to calibrate wheel sensors and nullify the accumulated errors. In [11], such solution is implemented for calibration and resetting the accumulated errors on public transport systems which use satellite and inertial navigation, but subway lacks access to satellite navigation.

Such approach has become worldwide [11–15]. The most similar solution is considered in [12], where passive tags are put on tunnel walls and readers are installed on head cars. The practice has primary focus on wide range of speeds, including 100 km/h. The authors propose using signal phase difference estimate received from a tag group (at least three in a group; total number depends on the speed) to locate trains. This method implies over-calculation, not every reader has the function of phase difference estimate. The authors have paid no attention to collision issues in conditions of multiple-tag scanning and cases when tags are missed due to high speeds.

From the author's point of view, RFID tag positioning of trains is more promising and simple in subways at the moment of first scanning of a tag by known session setup parameters of tag-reader communication. However, as seen from the experience of using a RFID-based positioning system in St. Petersburg Metro at relatively low speeds (less than 80 km/h) RFID tags are randomly missed. Thus, the need to study the issue of their improved detection became relevant.

Thus, navigation of trains on a line can involve wheel sensor readings of its traveled distance (or an equivalent device) completed by high-precision readings of a RFID-system.

Analytical review has shown poor coverage of the topic of reliability of RFID tag scanning at high speeds. Therefore, the first section covers features of readers of various generation and estimates reliability of tag reading at their high-

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

speed radio visibility zone. The second section studies accuracy of RFID-based positioning of trains.

II. Methods of Reliability Estimation of RFID Tag Scanning by Subway Trains

A. Review of Various Reader Generations and Their Operation Principles

Due to limited choice of solutions at the market, readers with process intervals in tag scanning were chosen for development of a non-contact track linking system in St. Petersburg Metro based on RFID technology.

Readers of this type require an external control signal to start which includes tag scanning time, and scanning time is selected within a limited range. After a start signal is received, the reader "searches" for a tag until one of the two possible options occur: Either a tag is successfully read, or no successful scanning until the scanning time expiration. At any result, further search is stopped and the results are transmitted to the car controller. Further scanning requires restarting the reader. Between the restart and the previous time expiration, there exist a certain interval. As the operating practices have shown this is the tag-reader communication algorithm, which features missing tag cases at allowed speeds, which reduced reliability of RFID-based positioning system operation.

Our study of this problem has led us to the conclusion that one of the ways to improve reliability of tag scanning is to shift to advanced readers with continuous scanning. Such readers require no external control signals to restart tag scanning: once a start signal is received, the reader scans non-stop until it receives a forced stop signal; tag detected does not stop scanning.

Procedure of train positioning against track coordinates in continuous scanning readers is design to be as follows:

1. A reader emits a continuous scanning signal, which is a cyclic sequence of requests to tags and wait for their responses. Each cycle lasts no longer than several milliseconds.

2. Each tag has a specific radio visibility zone, which depends on various parameters: strength of the signal emitted by the reader; attenuation in a reader-antenna path; antenna gain; parameters of the tag itself [16]. When the reader antenna appears within the radio visibility zone of a tag, reader-tag data communication becomes available. It is worth highlighting that the width of tag radio visibility zone is a random value normally distributed with average random of 2.3 m and root-mean-square deviation of 0.02 m, whereas radio visibility zone of the tag is symmetric to its location.

3. When receiving a request from the reader within its radio visibility zone, a tag responds within several milliseconds as per Gen2 protocol [7]. After having detected a tag, the reader initiates read out of the service data, including details on the tag location coordinates. Since the time spent on service data scanning from the tag does not equal to zero, when the reader antenna approaches the tag visibility zone at a specific speed, the latter will be scanned at a point further than its visibility zone acquisition one. It is apparent that the higher is the speed the further will be such point.

4. After having scanned the service data, the reader transmits the results to RFID-based positioning system car controller for further processing, after which the train is assigned a new coordinate. The time spent on processing and data exchange between the reader and the car controller is no longer than several milliseconds.

Introduce the term "tag scan time" and settle that it means a time interval from the moment when reader antenna enters tag visibility zone until the moment the car controller receives service data from the antenna.

According to bench tests, tag scan time is a random normally distributed value with average random of 45 ms and root-mean-square deviation of 5 ms.

Based on the above positioning algorithm for carriers of various generation readers against track coordinates, let us consider how accuracy and reliability of positioning is dependent from the speed, tag visibility zone variation and scan time. The solution search will involve mathematical modeling.

B. Methods to Estimate Reliability of RFID Tag Scanning by Subway Trains and Results of Scanning Procedure Modeling

Consider readers with external starting signal. Introduce a range of conventional notations and simplifications:

l – tag visibility zone width, m (within the regarded task taken as determined value equal to 2.3 m);

v – train speed when passing the tag visibility zone, m/s;

t_{del} – delay in reader control, s (random value);

t_{read} – duration of first powering up of tag and traffic, s (within the regarded task taken as determined value equal to 0.045 s);

t_{reread} – duration of tag repowering and radio traffic, s (random value);

t_{scan} – set scanning time, s.

Accept that delay in control over reader and the time of tag rereading are distributed with densities $f_{del}(t)$ and $f_{reread}(t)$, respectively.

With regard to the specified conditions, find possible results of passing tag visibility zone.

1. Tag will be read.

2. Tag reading will fail. Time for powering up of tag and traffic will exceed the time required to pass the tag visibility zone, i.e. $t_{read} > l / v$.

However, it is known that track tag visibility zone width is 2.3 m and the total time of powering up and scanning is 0.045 s. It follows that the tag will be missed at train speed of over 190 km/h due to the reason being considered, but such value cannot be obtained in Saint Petersburg subway. Thus, the specified result can be excluded.

3. Tag reading will fail: Train will come into the tag visibility zone with the controller disabled (due to search timeout during time $t \in [0, t_{del} - (l / v - t_{read})]$ and reaching the visibility zone), and after scanning restart will lack time to perform the scanning – no longer will be within the tag visibility zone.

4. Tag reading will fail. The reader will be disabled during tag powering up or radio traffic, i.e. at time interval $[0, t_{read}]$; and after the scanning field is recovered will lack time

to scan the tag before the train leaves the radio visibility zone.

Results 1, 3 and 4 will be assigned letters A, C_1 и C_2, respectively. It is apparent that the given results are incompatible and at the same time they create a complete group of events, i.e. $P(A + C_1 + C_2) = 1$. So, total possibility of missing a tag while passing its visibility zone can be calculated as:

$$P(\bar{A})(l, v, t_{scan}) = P(C_1) + P(C_2) \quad (1)$$

where $P(\bar{A}) = 1 - P(A)$.

Probability of result C_1 is conventionality and can be found using the formula:

$$P(C_1)(l, v, t_{scan}) =$$

$$\int_{\frac{l}{v}-t_{read}}^{\infty} f_{del}(t) F_{dis}\left(t + t_{read} - \frac{l}{v}\right) dt \quad (2)$$

where $F_{dis}(t)$ – probability of the reader leaving the scanning mode at time t; identical to uniform distribution function at the interval $[0, t_{scan}]$.

Similarly, probability of result C_2 can be found:

$$P(C_2)(l, v, t_{scan}) =$$

$$\int_0^{t_{read}} f_{dis}(t) \int_{\frac{l}{v}-t}^{\infty} (f_{reread} * f_{del})(\tau) d\tau \, dt \quad (3)$$

where $f_{dis}(t)$ – probability density function of disabling of the reader scanning field; identical to continuous uniform probability density function at the interval $[0, t_{scan}]$.

$*$ – convolution.

With consideration to (2) and (3) in the account, rewrite (1):

$$P(\bar{A})(l, v, t_{scan}) = \int_{\frac{l}{v}-t_{read}}^{\infty} f_{del}(t) F_{dis}\left(t + t_{read} - \frac{l}{v}\right) dt +$$

$$+ \int_0^{t_{read}} f_{dis}(t) \int_{\frac{l}{v}-t}^{\infty} (f_{reread} * f_{del})(\tau) d\tau \, dt \quad (4)$$

During continuous scanning tag scanning can fail in only one case, when the tag scanning time exceeds the time during which the train stays within the tag visibility zone. At speed of under 150 km/h and at scanning time of under 50 ms such event is impossible. Probability of reading tends to 1. *(Assumption: No hardware failures are considered.)*

C. Comparison of Readers as of Reliability of RFID Tag Scanning

Fig. 1 shows the results of mathematical model calculations for readers with external starting signal.

The acquired values are in line with observations at sections of subway line 4, according to which probability of missing tags subject to scanning times equal to one second at speeds over 60 km/h is $1.2 \cdot 10^{-4}$ (the result obtained from processing RFID-based positioning equipment logs).

If scanning time is increased to reach the maximum value of 65 seconds, the probability of missing tags at speed limits of St. Petersburg Metro is still significant $P > 1.2 \cdot 10^{-4}$.

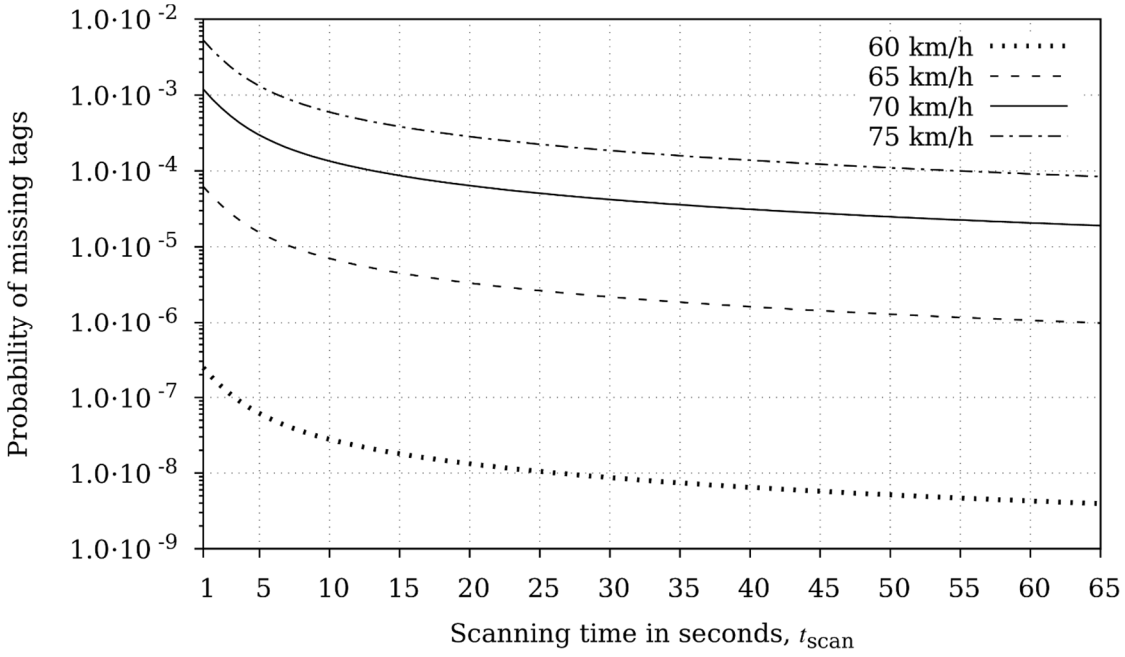


Fig. 1. Probability of missing tags depending on scanning time at various speeds.

D. Conclusions

1. Readers with external control of scanning feature limited reliability of RFID tag scanning.

2. Readers with continuous scanning feature very high reliability of RFID tag scanning (probability of tag scanning at speeds of max. 100 km/h tends to 1).

3. For accurate positioning at relatively high speeds it is reasonable to use continuous scanning readers.

III. Methods of Positioning Accuracy Estimation for Subway Trains with RFID Technology

A. Estimation Method for Train Positioning Accuracy with RFID Technology

Introduce the initial conditions:

1. Take tag location coordinates as known with zero error (these location coordinates are stored in the tag memory).

2. Regard a train as a mass point in the center of a RFID antenna.

3. Take the train speed as constant within a tag visibility zone and agree that its vector is directed to the track coordinate growth.

Introduce the notations:

l – tag visibility zone width, m (random value normally distributed with average random of 2.3 m and root-mean-square deviation of 0.02 m);

v – train speed when passing the tag visibility zone, m/s;

t_{read} – tag scan time, s (random value normally distributed with average random of 0.045 s and root-mean-square deviation of 0.005 s);

x_t – tag coordinate stored in its memory, m;

x_{read} – coordinate of the train at first tag scanning, m.

In such case, subject to the introduced values and notations, actual coordinate of the train at the moment of its positioning by tag can be calculated as function of speed.

$$x_{read}(v) = x_t - \frac{l}{2} + t_{read} v \tag{5}$$

where $x_t - l/2$ — the coordinate indicates tag visibility zone coverage, $l/2$ features symmetric character of its visibility zone.

Coordinate of the train at the tag scanning moment corresponds the accumulative of coordinates indicating tag visibility zone coverage and the distance traveled by the train during tag scanning time in relation to the coordinate indicating tag visibility zone coverage.

Define the deference between the train coordinate at the tag scanning moment of and the coordinates stored in the tag:

$$\Delta x(v) = x_t - x_{read}(v) = -\frac{l}{2} + t_{read} v \tag{6}$$

It is apparent that value $\Delta x(v)$ is random as it represents a result of summing up other two random values: $l/2$ – half-zone of tag visibility and $t_{read}\, v$, which is distance traveled by the train during tag scanning time. Where $\Delta x(v)$ is normally distributed as the sum of the two normally distributed random values shows normal distribution as well. To define parameters of the obtained distribution let us address the probability theory.

We know that multiplication of a random value x by a constant a will lead to mathematical expectation of the obtained random value $a\,x$ is found as:

$$M[a\,x] = a\,M[x] \tag{7}$$

Subject to expression (6) and definition of dispersion, we can get dispersion for random value $a\,x$:

$$D[a\,x] = M[(a\,x - M[a\,x])^2] = M[(a\,x - a\,M[x])^2] = M[\{a(x - M[x])\}^2] = M[a^2(x - M[x])^2] = a^2 M[(x - M[x])^2] = a^2 D[x] = a^2\sigma^2 \tag{8}$$

Thus, according to expressions (6), (7), and (8), distribution parameters for random value $\Delta x(v)$ can be calculated as follows:

$$\mu_{\Delta x}(v) = -\frac{\mu_l}{2} + v\,\mu_{t_{read}} \tag{9}$$

where μ_l – average random of the tag visibility zone width,

$\mu_{t\text{read}}$ – average random of the tag scanning time,

$$\sigma^2_{\Delta x(v)} = v^2\sigma^2_{t_{read}} + \frac{\sigma_l^2}{4} \tag{10}$$

where $\sigma^2_{t_{read}}$ – random dispersion of tag scanning time,

σ_l^2 – random dispersion of tag visibility zone width.

Let us consider the obtained results in more detail.

B. Evaluating Ultimate Accuracy of Train Positioning under Various Conditions

Through the above reviewed expressions, we have found family of densities Δx(v), which is difference between tag scanning coordinate and the coordinate stored in the tag memory according to the train speed within the tag visibility zone. Fig. 2 gives the results.

Mathematical expectation $\mu_{\Delta x}(v)$should be regarded as a bias in scanning which can be easily corrected at the point of train positioning. As seen from Fig. 2, the bias value reduces when the train speed grows, however, dispersion of distribution increases. Spread of values, characterized by dispersion $\sigma^2_{\Delta x(v)}$, against mathematical expectation defines the accuracy of RFID-based positioning of subway trains relative to track coordinates. Fig. 3 graphically represents dependence of train positioning bias and root-mean deviation of positioning error based on the speed at the specified parameters of initial distribution.

As seen from Fig. 3, positioning error at speeds close to zero stays within several centimeters. Such speeds are found within the head car stop at subway stations, where increased accuracy of stops is required subject to target braking. In case of cut-and-cover stations, positioning errors of several tens of centimeters are quite acceptable. For stations with platform screen doors, requirements are more strict and methods of increased accuracy of positioning should be developed.

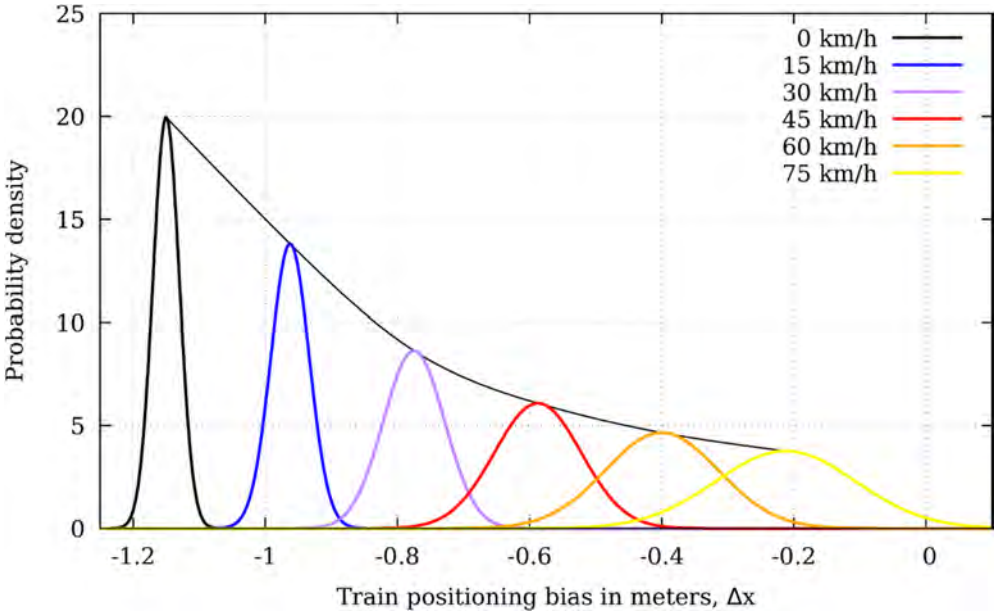


Fig. 2. Family of densities for random value Δx according to the train speed within the tag visibility zone.

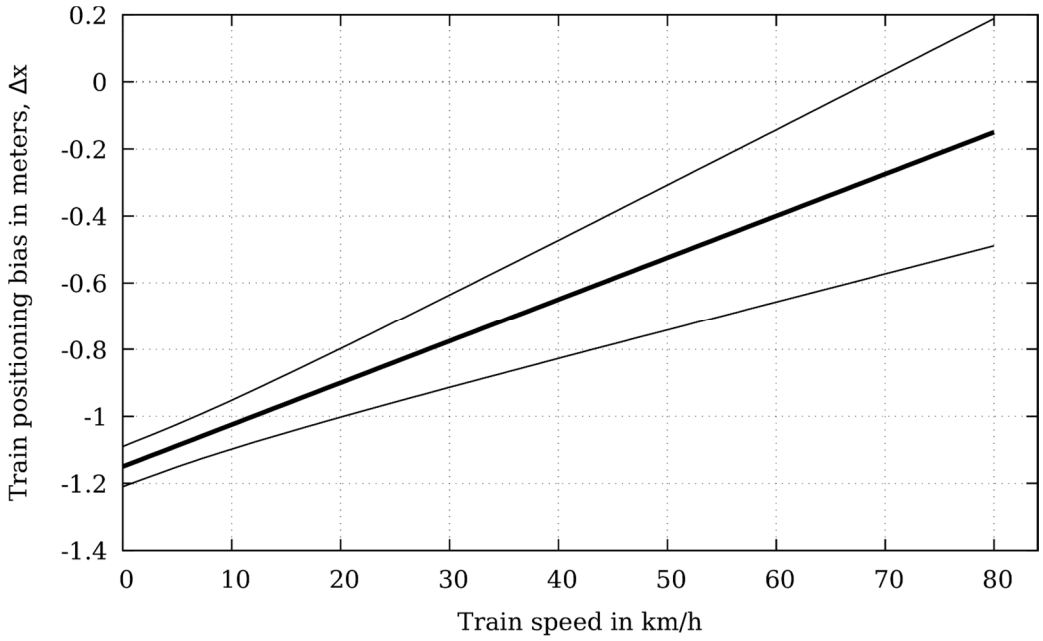


Fig. 3. Speed-dependence of Δx. Thick solid line designates the mathematical expectation of random value Δx. Fine lines designate upper and bottom boundaries ±3σ of taken estimates of Δx at a specified train speed.

At in-section speed of about 50 km/h, positioning error stays within ±20 cm. Such accuracy is sufficient for both automatic train operation, and high-speed diagnostics of linear objects.

C. Conclusions

1. Combination of a wheel odometer (installed on all vehicles) with RFID-equipment allows for highly accurate solution of the navigational task of train positioning as a necessary platform for any automatic train control system for trains.

2. The highest navigational accuracy is required at the stop zone at platform screen door stations, where slowdown speeds are low. Here, RFID-based positioning is capable of ensuring train positioning error of below ±10 cm, which is sufficient for target braking of trains at stations with platform screen doors.

IV. Conclusion

Fields of RFID application are growing wider. A very promising one appears to be RFID-based positioning. Which inevitably raises questions related to specific operating conditions of RFID tools. The use of RFID-based positioning for

solving railway transport tasks (automatic train control, linking high-speed diagnostic tool to tracks) has made research of improved reliability of RFID tag to ready communication topical, as well as research of improved accuracy of (train) reader antenna positioning against RFID tag locations.

Thus, this paper addresses an important issue of reliability of RFID tag scanning at various train speeds. The research results have shown that readers with external start signal feature low reliability of RFID tag scanning at relatively low speeds (tags can be missed) and require replacing for advanced continuous scanning readers in the fields requiring high reliability of tag scanning.

Continuous scanning readers allow for high accuracy of train positioning against RFID tag location. Expected accuracy of low-speed (below 10 km/h) positioning for the readers and tags reviewed in this paper is at least ±10 cm. Such accuracy is sufficient for target braking at stations with platform screen doors, which require the highest accuracy.

The result of the study are analysis methods that feature scientific novelty and allow for highly accurate solution of the navigational task of train positioning as a necessary platform for any automatic train control system for trains.

In particular, the results of the study can be used on railway transport and any other wheeled vehicles, which require high reliability and accuracy of positioning and independence from satellite navigation systems. Further author's research has it as the aim.

REFERENCES

[1] L. Baranov, and V. Maksimov "The Energy Efficiency of an Automatic Control System of Subway Train Movement and Requirements for its Subsystems", Russian Electrical Engineering, 2018. Vol. 89, Issue 9, pp. 546-549.

[2] L. Baranov, V. Maksimov, and N. Kuznetsov "Energy-Optimal Control of Vehicle Traffic", Russian Electrical Engineering, 2016, Vol. 87, Issue 9, pp. 498-504.

[3] J. Otegui, A. Bahillo, I. Lopetegi, and L. E. Díez "A Survey of Train Positioning Solutions", IEEE Sensors Journal, 2017, Vol. 17, Issue. 20, pp. 6788-6797, doi: 10.1109/JSEN.2017.2747137.

[4] A. Palmer, and N. Nourani-Vatani "Robust Odometry using Sensor Consensus Analysis", Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018, pp. 3167-3173, doi: 10.1109/IROS.2018.8594473.

[5] A. Kostrominov, E. Strekalovkskiy, and O. Tyulyandin "Analysis and increase of RFID-navigation accuracy in metro conditions" (in Russ.), Transport of the Ural, 2018, Vol. 59, Issue 4, pp. 23-27.

[6] D. Dobkin "The RF in RFID: Passive UHF RFID in Practice", USA, MA, Newton: Newnes, 2007, 504 p.

[7] EPCTM Radio-Frequency Identity Protocols Generation-2 UHF RFID, GS1 EPCglobal Inc. 2018.

[8] K. Finkerzeller "RFID Handbook. Fundamental and Applications in Contactless Smart Cards and Identification: 2nd edition", UK, Chichester: Willey, 2003, 427 p.

[9] A. Kostrominov, M. Korolev, V. Gavrilov, and T. Kryuchkova "RFID-technology application for automatic metro train operation" (in Russ), Proceedings of Petersburg Transport University, 2009, Issue 3, pp. 91-97.

[10] A. Kostrominov, and T. Kryuchkova "Operational algorithm for the traffic management system of metro electric stock based on RFID-technology" (in Russ.), Proceedings of Petersburg Transport University, 2014, Issue 2, pp. 42-48.

[11] Z. Wei, S. Ma, Z. Hua, H. Jia, and Z. Zhao "Train integrated positioning method based on GPS/INS/RFID", Proceedings of 35th Chinese Control Conference (CCC), Chengdu, 2016, pp. 5858-5862, doi: 10.1109/ChiCC.2016.7554274.

[12] A. Buffi, and P. Nepa "An RFID-based technique for train localization with passive tags", Proceedings of 2017 IEEE International Conference on RFID (RFID), Phoenix, AZ, 2017, pp. 155-160, doi: 10.1109/RFID.2017.7945602.

[13] M. Bouet, and A. L. dos Santos "RFID tags: Positioning principles and localization techniques", Proceedings of 1st IFIP Wireless Days, Dubai, 2008, pp. 1-5, doi: 10.1109/WD.2008.4812905.

[14] L. Xing, W. Meng, Q. Guangcheng, and L. Rong "LANDMARC with improved k-nearest algorithm for RFID location system", Proceedings of 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2016, pp. 2569-2572, doi: 10.1109/CompComm.2016.7925162.

[15] R. Wang, X. Ma, and Y. Wang "Application of improved RFID-based locating algorithm in locating of railway tunnel personnel", Journal of the China Railway Society, 2012, Vol. 10, Issue 34, pp. 68-71. 10.3969/j.issn.1001-8360.2012.10.011.

[16] A. Kostrominov, and T. Kryuchkova "Regression model of tags radio visibility zone for the system of non-contact binding to the underground track" (in Russ.), Transport of the Ural, 2012, Vol. 34, Issue 3, pp. 49-53.

Model and Means of Timed Automata-based Real-time Adaptive Transit Signal Control

Mykhailo Lytvynenko
Bachelor in Computer Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
mykhailo.lytvynenko1@nure.ua

Olexandr Shkil
Department of Design Automation
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
https://orcid.org/0000-0003-1071-3445

Inna Filippenko
Department of Design Automation
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
http://orcid.org/0000-0002-3584-2107

Leonid Rebezyuk
Department of System Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
https://orcid.org/0000-0001-8516-6584

***Abstract* — Mass transit systems are present in cities worldwide. To provide satisfactory level of surface public transportation quality and to prevent unnecessary injurious time competitions between high-capacity transit vehicles and private cars, transit signal priority techniques were considered to introduce in traffic light controller. Transit networks configurations were analyzed, and appropriate methods of transit vehicles motion management were introduced in the developed model of adaptive real-time transit signal control system.**

Keywords — transit signal priority, adaptive control model, timed automata, radio communication protocol, real-time control system

I. INTRODUCTION

Nowadays roads are becoming increasingly congested, counterintuitively construction of new ones does not solve a problem but creates an offer thus demand growth. In result it turns out a vicious circle where traffic jams do not disappear, and situation even becomes worse. Unfortunately, public transportation efficiency is gravely affected by situation caused by this vicious circle. In the same traffic jam may be blocked a car carrying in average 1.2 -1.7 persons and a tram or a bus whose capacity is at least 10 times more and may extend to several hundred passengers. Car drivers do not see any reason why they should move from their own car to public transportation services and furthermore such public transport conditions force new people to choose a private car to commute thus aggravating traffic.

Or another situation: on intersection, tram must wait for the permissive traffic signal phase. Supposing that it transports 80 passengers and delayed for half-minute while conflicting direction traffic flow does not exceed 30 vehicles/min, so one vehicle every 2 seconds, person delay for tram is 40 minutes which is equivalent for car with two passengers to be delayed for 20 minutes at the intersection! While for other vehicles person delay is around 8 seconds considering that in optimal applications just around 10 seconds is enough for tram to traverse a regular intersection if it was detected in advance, so no time lost during acceleration.

Finally, it is common when bus or tram does not have enough time to pass the intersection if it is in traffic jam on shared lane before the traffic signal. After permissive phase enabled, transit vehicle must wait to proceed after prior vehicles and may catch a prohibitive phase of traffic light when transit vehicle reaches it. In result, vehicle loses approximately a half of signal cycle duration.

Described cases could relatively easily be improved using dedicated public transportation lanes (i.e. space priority) alongside with various active transit signal priority (TSP) methods to control transit and traffic flows whose objective is to reduce delays of public transportation vehicles with ideally the less possible impact on other vehicles traffic (i.e. time priority). Active method [1] of TSP providing utilizes interaction between transit vehicle and infrastructure, contrarily to passive method [2] relying on statistical data only about transit route or network in general (e.g. schedule, dwell time etc.). Active priority methods range from the simplest ones where only check-in vehicle detector is necessary to more complicated with GPS or AVL (automatic vehicle location) systems being involved.

Currently, the problem of providing transit space priority is mostly solved but it remains the problem of providing transit priority on signalized intersections by timely traffic light switching. Alternative solutions can not provide interaction with distinctive Ukrainian transit signals (Fig. 1) without expensive traffic light devices replacement or adaptation of existing control approaches. So, the goal and objective of this research is the development of an adaptive transit signal control model as well as model-based real-time control system implementation to improve surface transit circulation performance. Scientific novelty is a timed automata-based model of an adaptive real-time transit signal control to minimize person delay on signalized intersections.

At present, in the city of Odesa, Ukraine a TSP system is operated [3] where a countdown timer is called at detection of any vehicle, after timer elapsed, vehicle is provided with permissive phase. However, such system does not provide transit vehicle route identification which is important on intersections with several directions-to-go and for deep priority request analysis.

II. ADAPTIVE REAL-TIME TRANSIT SIGNAL CONTROL MODEL

A. Model scope

The model is developed considering that based on this model real-time system will control Ukrainian standardized

T-shaped transit signal (Fig. 1) presently being used for tram circulation control only.

Fig. 1. T-shaped transit signal with permissive straightforward phase enabled

The model considers a separate transit signal only and its reaction on approaching public transportation vehicles of specific routes. Since control signals for mass transit and other traffic members are often different, so it is mandatory to use different types of traffic lights to avoid any confusion.

B. Transit signal operational modes

Here are presented possible states of T-shaped transit signal used in tram systems across Ukraine and how they are split in operational modes of transit signal inside the model (Fig. 2) so approaching of vehicle of specific direction route will unambiguously define the state of the model after transition.

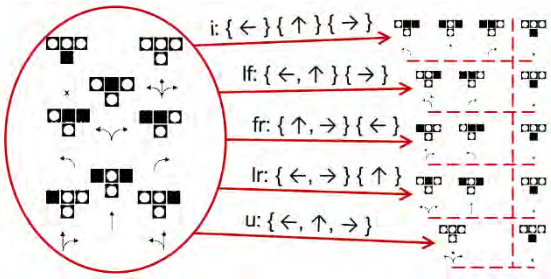


Fig. 2. Possible transit signal states and operational modes definition

Since it is not reasonable to allow movement in any direction by detection of vehicle that is supposed to happen in u (universal) mode, so it is set to be generally time-driven. Unlike universal mode, in other modes transitions into every state which is in fact transit signal phase are called by approaching vehicle.

C. Means of providing a priority

The model implements mainly transit "phase insertion" approach of TSP and optionally "green extension" (Fig. 3) also, which is particularly useful in universal mode of transit signal operation or in other modes in case of very busy locations with many frequent transit lines to serve. Approach of "green extension" prevents the last of two following vehicles to be delayed at the intersection if they need to proceed in the direction already allowed by permissive phase enabled by previous vehicle. Still in some cases it might cause an unwanted vehicle bunching but it is unsolvable issue without reliable schedule information. In universal mode it is less problematic since basic signal phases are time-driven, here "green extension" just allow to vehicle to pass an intersection even if it arrives at the end of prearranged permissive phase, it will be provided with some additional seconds to prevent waiting for the next traffic light cycle.

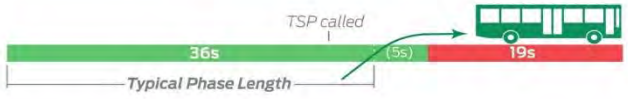


Fig. 3. Operation of TSP in "green extension" mode [1]

Adaptive real-time transit signal control model is developed using timed automata approach [4]. Timed automata are an extension of finite state automata with a finite number (but arbitrary) clocks in continuous time. The state diagram of control timed automata is given on Fig. 4.

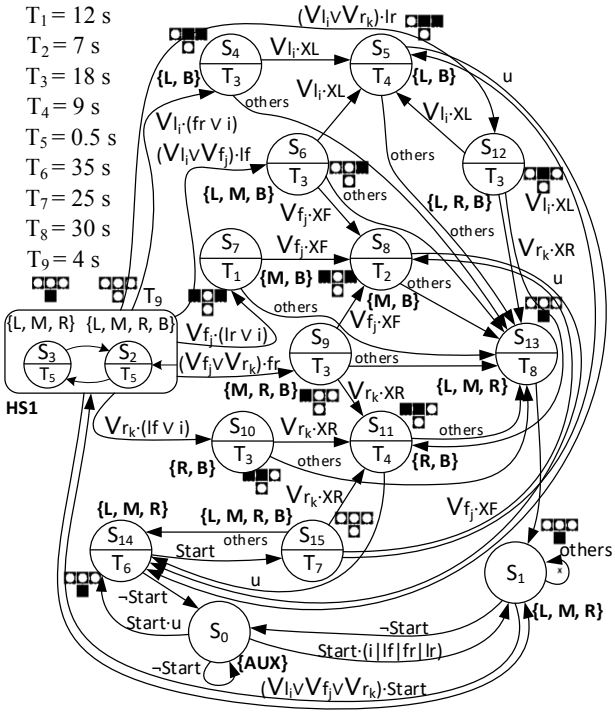


Fig. 4. Timed automaton state diagram of adaptive control model

All states except 0 and 1 are temporal meaning that process will not leave such state until time T_i is elapsed. States inside hyper state HS1 are time-driven only and do not react on external conditions until outer timer is elapsed. Transition conditions details will be described further in the paper.

III. CONTROL SYSTEM CAPABILITIES

The system developmental prototype based on described model would allow to optimize mass transit vehicles circulation according to certain criteria (reduction of delays at intersections as well as the amount of fleet required to serve public transportation network lines, increase the regularity of circulation) using data from such vehicles obtained in real time. It consists of two main parts: vehicle model (Fig. 5(1-2)) and transit signal control system model (Fig. 5(3-5)). They both are equipped with radio modules allowing them to communicate with each other. Technology of radiocommunication used in the system is Bluetooth. Due to the nature of vehicle to infrastructure communication implemented in this system, master device is the vehicle

model and slave is the transit signal controller. However, communication performs in duplex mode.

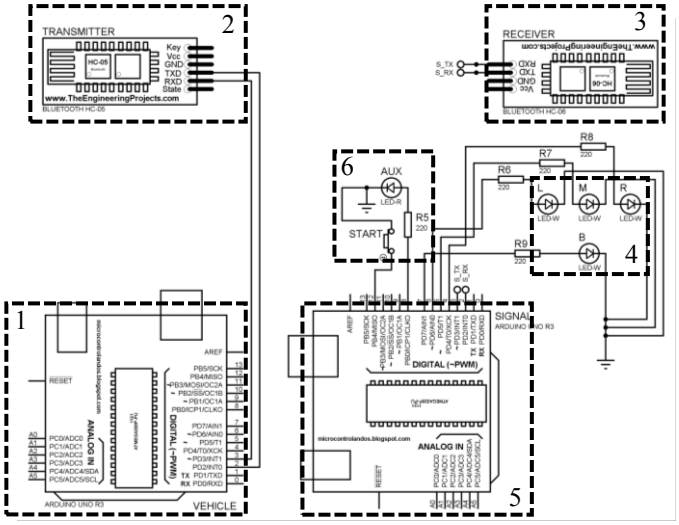


Fig. 5. Wiring diagram of the system components: (1) Vehicular intelligent transportation system model; (2) Vehicular radio module; (3) Signal controller's radio module; (4) Indication block; (5) Transit signal controller model; (6) Service indicator and sensor

A. *Servicing discipline*

The system is real-time with firm requirements [5] so request as detection of radio signal from transit vehicle should be processed within a few minutes, otherwise system will operate improperly with certain likelihood. However, in very busy parts of transit network the real-time system is supposed to operate with hard requirements so signal handling from transit vehicle must be accomplished within a few seconds before arrival of following vehicle or system will definitely fall.

Actually, processing of wireless signal from vehicle does not necessarily mean that this vehicle will be immediately allowed to pass the intersection by enabling permissive transit signal phase for it. Handling of this signal just supposes adding related transit line to waiting FIFO queue. Every item in this queue is served as soon as possible according to internal state of the control model. The waiting queue never overflow meaning that system always remains in the stationary mode. This is achieved by reserving enough space for storing information about waiting transit as well as by physical limitations (Fig. 6) of short-range radio communication used in the system [6].

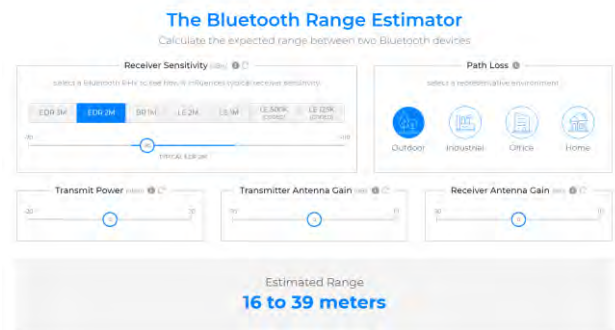


Fig. 6. Communication range estimation [7] according to equipment parameters and operating conditions

B. *Line information and its representation*

Vehicular signal carries information about line number, origin, destination terminuses and vehicle's numerical ID. This basic info is sufficient to provide clear identification of the detected vehicle and which route it is following. For example, using such identification the system will not react on vehicles of opposite direction as well as it will be able to distinguish two consecutive vehicles even of the same route. This info, except ID, is supposed to be entered in the system by tram/bus driver before the departure from depot. During the ride entered information will be transmitted in form of messages.

Here is explained how info about line is stored inside vehicular and transit signal parts respectively: *{{NNN<<PPP>>DDD^^Dv-...Dv-##VEHNUM}* and *NNN<<PPP>>DDD^^Dv-..Dv-#\n* where NNN is line number (2 bytes), PPP and DDD are provenance and destination codes (2 bytes), Dv is a deviation code (1 byte), there may be a list of deviations. Finally, VEHNUM represents vehicle registration number within public transportation facility (4 bytes). This component is not known during the transit signal controller setup, but such field allows to prevent controller from constant analysis of the signal from one vehicle.

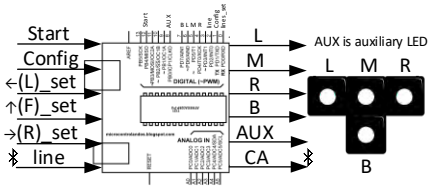
As mentioned previously, system allows to treat differently routes with same terminuses but different intermediate stops. Such cases are called deviations. They may occur both in regular network operation and in case of accidents that change network configuration and vehicles have to follow corrected routes. Deviations handling concerns only transit signals where routes are branching. In other locations it is sufficient to provide expected direction-to-go for this line regardless if it is deviation route or regular one.

IV. System Operation Algorithm

The algorithm of the adaptive real-time transit signal control system operation provides the following procedures.

A. *Setup part*

Operation of the system starts with setup definition which at the current stage is performed via USB-interface by changing configuration parameters (Fig. 7) in firmware source code. Configurable are time parameters (durations) of phases, activation of phase extensions and their directions, mode of signal operation, deviations handling. Also, the system needs to be provided with expected transit lines and their directions-to-go. These lines are split into three sets (left, straight and right) and in case of deviations handling enabled both regular and deviation route directions must be specified.



Config = {mode, ext. left, ext. forward, ext. right, deviations}

Fig. 7. Inputs and outputs of the transit signal control system

On the figure "line" input refers to currently detected line and "Start" is a signal from launch button. Outputs all except one are signal indicators, auxiliary indicator utilized to

confirm the system serviceability in its idle state, CA output represents reverse communication from signal to vehicle to make sure intersection clearance by the vehicle being currently served.

B. *Vehicle serviving procedure*

After configuring the system and providing information about expected transit routes, transit signal prototype is ready to serve approaching vehicles. After pressing start button, depending on whether the system is set to u (universal) mode, after idle state 0 (Fig. 4) timed automaton either goes to temporal state 14 or to standby state 1. Further transition from it will trigger only on signal appearing from vehicle of any of the intended routes. Expression $\vee x_i$ means, that transition will trigger only in case of the vehicle detection of route x_i from the direction set X. It works like elementwise OR operation.

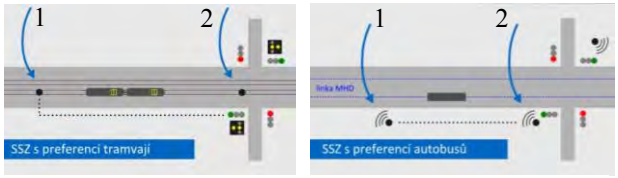
Hyper-state HS1 implements blinking of the bottom transit signal indicator, informing the vehicle's driver that the priority phase will be provided soon. Then, in accordance with the direction-to-go of the route and the operational mode of the system, the transit "phase insertion" is performed by triggering transition to state 4, 6, 7, 9, 10 or 12. After elapsing the time allotted for vehicle servicing procedure, the system returns to timeout state 13, or in case of phase extension enabled, and, if there is a signal from the next vehicle, also gives it permission to traverse the intersection. Phase extension in a certain direction can occur only if the previous phase already permitted movement in this direction.

C. *Clearance acknowledgement considerations*

At the transition from permissive phase top vehicle of the waiting queue is not just popped out but before the leaving of current state the model performs so-called "clearance acknowledgement". This means that the system makes sure that recently served vehicle has well crossed the intersection and its signal no longer detected by the system. Otherwise it might cause huge problems since if for some reasons vehicle was blocked at the intersection and did not leave it, the following vehicle will encounter the same problem. And finally, from the models's point of view there is no waiting public transport vehicles, while in reality there are two blocked vehicles which cannot proceed even after resolving problems that were caused the delay because the system is not supposed to give the permission to pass an intersection.

Since there is no direct way to check that the vehicle to infrastructure connection is broken and just such check would have insufficient reliability, to overcome this issue the request-response technique was utilized. The idea is to check the presence of current vehicle by sending a special message and comparing the conditionally received message's data with initial one. If it does not match or there is more than one vehicle in the queue, so previous vehicle is considered to clear an intersection since because of the nature of Bluetooth communication only one link between master and particular slave can be established at the given moment of time.

Important to note that check-in detector (Fig. 8(1)) must be installed a few hundred meters ahead [8] of dynamically controlled traffic light and check-out detector (Fig. 8(2)) should be installed near the transit signal.



a) b)

Fig. 8. TSP detectors location near intersection for tram (a) and bus (b): (1) check-in detector; (2) check-out detector with clearance acknowledgement option

But at the current stage they are combined for simplicity of system testing and modeling.

V. CONCLUSION

In this manner, developed adaptive control system allows real-time transit signal switching respectively with transit vehicle approaching. This minimizes vehicle downtime thus reduces current vehicle energy consumption and improves general transit service regularity. Advantages of the system in comparison with analogues are downtime reduction to a minimum and transit route identification capability. It is planned to perform an experimental study to determine the optimal delay of transit signal switching time as well as to integrate this system with existing cloud traffic control system [9].

REFERENCES

[1] "Active Transit Signal Priority | National Association of City Transportation Officials", National Association of City Transportation Officials, 2016. [Online]. Available: https://nacto.org/publication/transit-street-design-guide/intersections/signals-operations/active-transit-signal-priority/. [Accessed: 26- Jul- 2020].

[2] "Transit Signal Progression | National Association of City Transportation Officials", National Association of City Transportation Officials, 2016. [Online]. Available: https://nacto.org/publication/transit-street-design-guide/intersections/signals-operations/transit-signal-progression/. [Accessed: 26- Jul- 2020].

[3] "Public transport priority was implemented in Odessa", Komkon.ua. [Online]. Available: http://komkon.ua/en/prioritetnyj-proezd-obshhestvennogo-transporta-byl-vnedrjon-v-odesse/. [Accessed: 31- Jul- 2020].

[4] S. Guellati, I. Kitouni, R. Matmat and D. Saidouni, "Timed Automata with Action Durations – From Theory to Implementation", Communications in Computer and Information Science, vol. 465, pp. 94-109, 2014. Available: 10.1007/978-3-319-11958-8_8 [Accessed 31 July 2020].

[5] T. Kaldewey, C. Lin and S. Brandt, Firm Real-Time Processing in an Integrated Real-Time System. Santa Cruz: Computer Science Department, 2006.

[6] "HC-05 Bluetooth Module Pinout, Specifications, Default Settings, Replacements & Datasheet", Components101.com. [Online]. Available: https://components101.com/wireless/hc-05-bluetooth-module/. [Accessed: 26- Jul- 2020].

[7] "Understanding Bluetooth Range | Bluetooth® Technology Website", Bluetooth® Technology Website. [Online]. Available: https://www.bluetooth.com/learn-about-bluetooth/bluetooth-technology/range/. [Accessed: 26- Jul- 2020].

[8] Y. Lin, X. Yang, N. Zou and M. Franz, "Transit signal priority control at signalized intersections: a comprehensive review", Transportation Letters, vol. 7, no. 3, pp. 168-180, 2014. Available: 10.1179/1942787514y.0000000044 [Accessed 26 July 2020].

[9] V. Hahanov, A. Ziarmand and S. Chumachenko, "Transportation Computing: "Cloud Traffic Control"", Cyber Physical Computing for IoT-driven Services, pp. 201-217, 2018. Available: 10.1007/978-3-319-54825-8_10 [Accessed 3 September 2020].

Geometry-Based Rolling-Stock Identification System Insensitive to Speed Variations

Valery A. Zasov
Samara State Transport University

Samara, Russia
vzasov@mail.ru

Maxim V. Romkin
Samara Center for Diagnosis and Monitoring of Infrastructure Facilities
Samara, Russia
romkinmaks@rambler.ru

***Abstract*—This paper proposes a geometry-based rolling-stock identification system that reliably identifies types of railroad vehicles when the train is traveling at a varying speed. A feature of the proposed system is that it adjusts the measured distance from the axle of the last wheelset of a railroad vehicle to the axle of the first wheelset of the next railroad vehicle depending on the magnitude and sign of acceleration. This compensates for measurement errors that occur when the train is traveling at a varying speed. This paper also describes the identification system's algorithm and proposes potential applications for the system.**

Keywords—railroad transport, train, rolling stock, identification system, geometry, nonuniform movement, measurement adjustment, operation algorithm

I. Introduction

The efficiency and safety of rolling-stock operation largely depend on how complete and accurate information on the travel of locomotives and railcars is. Contemporary identification systems operate on various physical principles [1].

The widespread method of axle counters [2] and the method of track circuits [3,4] do not allow determining the types of rolling stock, therefore, they have limited functionality for solving identification problems.

Rolling-stock movement is monitored with identification systems based on the principle of scanning information from train-mounted sensors or tags.

Examples include optical character recognition (OCR) systems such as COMBAT [5], ARSCIS [6], which identify car registration numbers; radio-frequency identification (RFID) systems [7] such as Palma (Russia) [8], Amtech (the USA), and Dynicom (Europe) [9], based on the radio-frequency scanning of on-board encoders; and systems based on satellite navigation systems such as GPS [10] and GLONASS (KLUB-U).

A feature of systems currently in use is having to outfit vehicles with identification sensors or tags. This makes identification less reliable when the climatic or weather conditions are harsh, car numbers are soiled, or on-board sensors are damaged during loading, unloading, shunting, and suchlike operations.

Since they are complicated and costly, scanners for rolling-stock identification systems are not used except at major junction stations.

Reference [11] propose a system for identifying rolling-stock types by measuring and monitoring an aggregate of their design parameters, such as the number of axles, distances between the axles, overall coupler-to-coupler length, and other linear dimensions. These structural and chiefly geometric parameters are naturally inherent in all types of railroad vehicles and do not require outfitting railroad vehicles with specialized equipment.

Rolling-stock monitoring systems (RSMSs) [12] widely operated by railroad stations across Russia can be used for geometric measurements and identification. RSMS equipment can detect overheated axle boxes, wheelsets with defects and sticking brakes, and overloaded cars while the train is moving [12].

RSMSs have almost all the necessary tools to determine the geometric parameters used as input data in the geometry-based identification algorithm [11].

This algorithm can be implemented as software based on modules that complement the RSMS software, expanding its functionality to solve the problem of identifying rolling stock.

The geometry-based identification system [11] is by far simpler and cheaper to operate, but it identifies only the types and subtypes of locomotives and freight and passenger cars. Therefore, it is advisable to use this system to supplement the OCR [5,6] and RF [7,8,9] ones operated at major junction stations. With a geometry-based identification system in place, RSMS-equipped intermediate stations can provide total, end-to-end monitoring of rolling-stock movement.

This makes the operational management of rolling stock more efficient and allows you to control the train schemes in the process of movement for compliance with applicable regulations.

A major cause affecting the reliability and accuracy of geometry-based identification in known system [11] is the significant error that occurs in parameter measurements when the train is moving at a varying speed (accelerating or slowing down). One example of measured parameters is the overall coupler-to-coupler length. If determined incorrectly, this length may lead to errors or uncertainty in identifying the type of railroad vehicle.

Therefore, developing a geometry-based identification system that is not affected by speed variations is a relevant problem. This paper concerns itself with solving it.

This paper proposes a geometry-based rolling-stock identification system that reliably identifies types of railroad vehicles when the train is traveling at a varying speed.

II. Selecting and Justifying Geometric Parameters for Rolling-Stock Identification

To identify types of rolling stock, we propose applying a set of geometric parameters—the inter-axle distances shown in Fig. 1, which fall into three classes, $M^{(v)}$, $M^{(t)}$, and $M^{(go)}$:

1. $m_k^v \in M^{(v)}$, the inter-axle distance between the last wheelset of the first bogie and the first wheelset of the second bogie

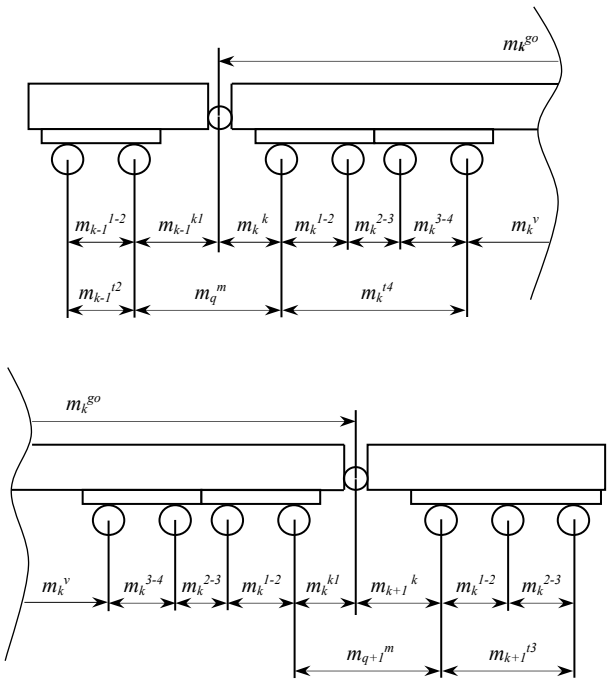


Fig. 1. Inter-axle distances and their relative positions

2. $m_k^{ti} \in M^{(t)}$, the distance(s) between the bogie axles. For a four-axle (i = 4), a three-axle (i = 3), and a two-axle (i = 2) bogie, these distances are, respectively:

$$m_k^{t4} = m_k^{1-2} + m_k^{2-3} + m_k^{3-4},$$

$$m_k^{t3} = m_k^{1-2} + m_k^{2-3},$$

$$m_k^{t2} = m_k^{1-2},$$

where $m_k^{1-2}, m_k^{2-3}, \text{and } m_k^{3-4}$ are the distances between the axles of the first and second, the second and third, and the third and fourth wheelsets in the bogie.

3. $m_k^{go} \in M^{(go)}$ is the overall coupler-to-coupler length of the vehicle, and it is calculated from

$$m_k^{go} = m_k^{k} + m_k^{ti} + m_k^{v} + m_k^{ti} + m_k^{kl},$$

where m_k^{k} and m_k^{kl} are the distances from the coupler axes to the axles of the outermost wheelsets at the vehicle's ends.

Thus, we propose identifying types of railroad vehicles, e_k, by determining the following group of geometric parameters (inter-axle distances) [11]:

$$e_k = \left\{ m_k^{go}, m_k^{ti}, m_k^{v} \right\}$$

Our system analysis of engineering and reference literature and our modeling and field experiments confirmed the validity of the proposed approach and the possibility of unambiguously identifying various railroad vehicles.

We propose identifying railroad vehicles at three hierarchical levels: by type, by subtype, and by model.

According to their type, railroad vehicles are classified into locomotives and cars. Subtypes of locomotives include electric locomotives and diesel locomotives; subtypes of cars, passenger and freight cars. Electric and diesel locomotives are further classified by model (e.g., the 2ES5K mainline electric locomotive), and so are passenger and freight cars. Freight car models, for example, include covered cars, gondola cars, flatcars, hopper cars, tanks, dump cars, and transporters.

A major factor affecting the reliability of geometry-based identification is the variation in the inter-axle distances used to describe the types of railroad vehicles.

The inter-axle distances used to identify the types of railroad vehicles are classifiable into two groups.

The first group includes inter-axle distances determined by the design and manufacturing process of the railroad vehicle. These values are almost constant (with a spread not exceeding 10 mm) and do not depend on whether the train is moving or stationary. An example is the distance m_k^{ti} between the bogie axles.

The second group includes inter-axle distances that depend on the magnitude of acceleration when the vehicle is moving. An example is the overall coupler-to-coupler length m_k^{go}. The overall length includes the distances m_k^{k} and m_k^{kl} formed by the mounted parts on vehicles with automatic couplers, containing elastic components to dampen shock loads from traction and braking forces.

The coupler dampers (spring, friction, or rubber–metal ones) cause the inter-axle distance to change: during braking, it decreases; during acceleration, it increases. During braking and acceleration (traction), the distance varies for different types of cars between 75 and 110 mm [13].

The overall length m_k^{go} is an attribute used to identify the type of railroad vehicle, and its stationary-state value can be found in the specifications for the vehicle. If determined incorrectly when the train is moving at a varying speed (accelerating or slowing down), this distance may lead to errors or uncertainty in identifying the type of railroad vehicle.

To make identification more reliable and accurate under varying speed conditions, we propose adjusting the measured overall coupler-to-coupler length m_k^{go}.

The adjustment is based on the measured magnitude and sign of acceleration for the train and takes place automatically [13].

The signal-separation methods described in [14,15] are used to calculate inter-axle distances with the distorting effects of interference and uneven movement eliminated. This makes rolling-stock identification reliable and accurate when the train is moving at a nonuniform speed.

III. Algorithm for the Rolling-Stock Identification System

The inter-axle distance $e_k = \left\{ m_k^{go}, m_k^{ti}, m_k^{v} \right\}$ is determined by measuring the vehicle's speed and the time the wheelsets take to cover track segments equal to the inter-axle distance. Fig. 2 (*a–d*) explains the principle of these measurements.

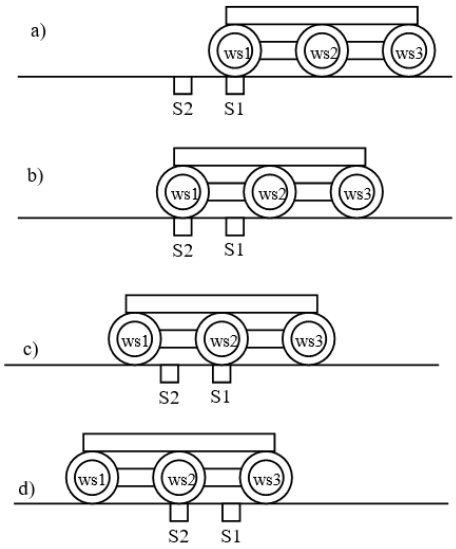


Fig. 2. Determining the inter-axle distance

For measurements, inductive sensors [6] S1 and S2 are used that register wheelset (WS) travel. The sensors are spaced one meter apart, which is less than the smallest inter-axle distance of railroad vehicles in use. The instantaneous speed is determined by measuring the time taken by wheelset WS1 to cover the distance between sensors S1 and S2 (Fig. 2-a and Fig. 2-b).

Then the time taken by wheelsets WS1 and WS2 to pass sensor S1 (Fig. 2-c and Fig. 2-d) is measured, and the inter-axle distance m_k^{1-2} is calculated. The other inter-axle distances are determined similarly.

These measurements and calculations take place after the identification system is turned on by a signal from the proximity sensor that registers the entry of a locomotive into the monitoring zone.

Once the inter-axle distance is determined between the first and second (m_k^{1-2}) and the second and third (m_k^{2-3}) axles of the bogie, its model and specifications are identified. The specifications are the railroad vehicle's number of axles and type (locomotive or car).

Next, with the train in motion, the inter-axle distance m_k^{v} is determined from the last wheelset of the first bogie to the first wheelset of the second bogie. This distance and the bogies' inter-axle distances allow the subtype of railroad vehicle to be identified.

If the vehicle is a locomotive, its subtype (diesel or electric) and commercial model are identified, with the overall coupler-to-coupler length m_k^{go} included in the model's specifications. If the vehicle is a car, the system identifies whether it is a passenger or freight one.

Determining the model of a car requires finding its overall coupler-to-coupler length m_k^{go}. To determine the length m_k^{go}, sensors S1 and S2 measure the distance m_q^{m} between the axle of the last wheelset of the preceding vehicle and the axle of the first wheelset of the next vehicle (Fig. 1).

The first vehicle in a moving train is the locomotive; its type and model are identified in the manner described above. From the specifications for the locomotive model, its overall dimensions are known, including the length of its mounted component m_{k-l}^{kl}. Then the length m_k^{k} of the mounted component on the vehicle following the locomotive (the second locomotive unit, the second locomotive, or a car) can be determined as follows (Fig. 1):

$$m_k^{k} = m_q^{m} + m_{k-l}^{kl}.$$

The lengths of mounted components at the vehicle's ends are equal: $m_k^{k} = m_k^{kl}$. From this it follows that the overall coupler-to-coupler length m_k^{go} is given by

$$m_k^{go} = 2m_k^{k} + m_k^{ti} + m_k^{v} + m_k^{ti}.$$

Thus, after determining the overall length and the length of mounted components for the vehicle, one can determine those for the next vehicle, and so on. This allows freight and passenger car models to be identified accurately.

Because of the coupler dampers, the distance changes between the axle of the last wheelset of the preceding vehicle and the axle of the first wheelset of the next vehicle when the train is braking or accelerating. In these cases, the measured inter-axle distance m_q^{m} is not equal to any reference value in specifications for the various types of railroad vehicles. This affects identification accuracy, resulting in errors and uncertainty.

To improve the reliability of identification under varying speed conditions, we propose adjusting the measured distance m_q^{m} from the axle of the last wheelset of a railroad vehicle to the axle of the first wheelset of the next railroad vehicle depending on the magnitude and sign of acceleration.

This compensates for the errors that occur in calculating the vehicle's overall coupler-to-coupler length m_k^{go} when the train is moving at a nonuniform speed [13].

Implementing the adjustment algorithm involves saving to the system memory all possible values of the inter-axle distance m_q^m sourced from specifications for the railroad fleet. Acceleration is measured by calculating the difference in speed between two adjacent wheelsets. The acceleration so obtained is then compared with the threshold value. An acceleration exceeding the threshold has an effect on the variation of the inter-axle distance m_q^m.

The principle for adjusting inter-axle distances is shown in Fig. 3, where $m_{q(i+1)}^m > m_{qi}^m > m_{q(i-1)}^m$ are the reference values of the inter-axle distances for various coupled vehicles, and the values Δ_{i+1}, Δ_i, and Δ_{i-1} indicate the ranges in which those distances vary when the train is accelerating or braking.

If the measured acceleration is less than the threshold, a signal is generated indicating the absence of acceleration—that is, the uniform movement of the train. This condition is marked by point A in Fig. 3.

In this case, the inter-axle distance is not adjusted, and an associated value is retrieved from the system memory whose difference from the measured value is minimal in modulus: m_q^m.

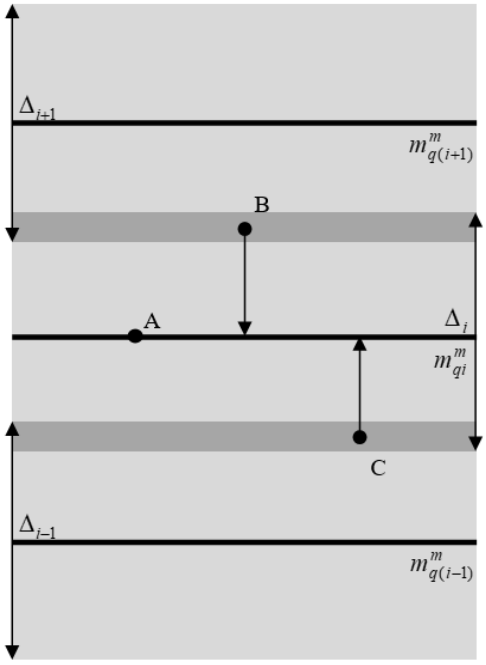


Fig. 3. The principle for adjusting the inter-axle distance m_q^m under various acceleration conditions

If the measured acceleration is greater than the threshold, two signals are generated: one indicates a positive acceleration (the train gaining speed); the other, a negative acceleration (the train slowing down). These signals retrieve from the system memory two values for the inter-axle distance that have a minimum difference from the measured value m_{qi}^m, both with the plus sign (distance $m_{q(i+1)}^m$) and with the minus sign (distance $m_{q(i-1)}^m$).

The magnitude of the inter-axle distance during acceleration increases above the reference value m_{qi}^m. Indeed, with a positive acceleration, the coupler dampers are stretched. This condition is marked by point B in Fig. 3. In this case, the smaller of the inter-axle distances m_{qi}^m and $m_{q(i+1)}^m$ from the memory (i.e., the distance m_{qi}^m) is taken as the adjusted inter-axle distance.

When the train is braking, the inter-axle distance diminishes compared with the reference value m_{qi}^m because with a negative acceleration, the coupler dampers are compressed.

This condition is marked by point C in Fig. 3. In this case, the greater of the inter-axle distances m_{qi}^m and $m_{q(i-1)}^m$ from the memory (i.e., the distance m_{qi}^m) is taken as the adjusted inter-axle distance.

Then the above steps are completed to calculate the overall coupler-to-coupler length m_k^{go}, and the vehicle's model is identified.

Once each vehicle passes sensors S1 and S2, an entry is generated in a dedicated file entitled Train Formation to list the identification results. The entry includes the type, subtype, model, and serial number of the railroad vehicle. The serial number serves as the entry's ID.

When the last car passes the monitoring zone, the Train Formation file is completed, and it is assigned an identifier containing its name and the date and time the file was created. The train's last car is identified from comparing the inter-axle distances m_q^m between the cars with a given threshold value. If the value m_q^m exceeds the threshold, the identified car is not followed by another, meaning the car is the train's last.

Table 1 outlines the above sequence of steps as a generalized algorithm.

Geometry-based identification systems are mounted at the boundaries of railroad stations in locations equipped with RSMS devices. For that reason, Train Formation files are generated both when the train arrives at and when it departs from the station. These files are sent to transportation control centers, making it possible to monitor train integrity when the train is traveling between stations as well as changes to train formations during loading, unloading, shunting, and other operations at stations.

A feature of the proposed algorithm is that it presents the solution to a complex multidimensional identification problem as sequential solutions to smaller identification problems. Identification consists in determining the type, subtype, and model of the railroad vehicle (in that order).

This feature reduces the performance requirements for the system's computer.

TABLE I. GEOMETRY-BASED IDENTIFICATION ALGORITHM

Item no.	Step
1	Check for a signal from the train proximity sensor. Signal present: the system turns on and moves to step 2. No signal: the system remains in standby mode
2	Identify the type of railroad vehicle (locomotive or car). Locomotive: proceed to step 3. Car: proceed to step 4
3	Identify the subtype of locomotive (electric or diesel). Identify the mass-produced model of the electric or diesel locomotive. Proceed to step 7
4	Identify the subtype of car (passenger or freight). Proceed to step 5
5	Determine the car's overall coupler-to-coupler length. Adjust the length according to the magnitude and sign of acceleration. Proceed to step 6
6	Identify the car model. Proceed to step 7
7	Save the identification results to the Train Formation file. Proceed to step 8
8	Check whether the vehicle is the train's last. If so, proceed to step 9; otherwise, return to step 2
9	Complete the Train Formation file, send it to the transportation control center, and return to step 1

Each of the train's vehicles is identified right after the last one passes the monitoring zone. Once the last car passes the monitoring zone, the Train Formation file is completed and is ready to be processed at a transportation control center.

IV. MODELING RESULTS

Two models were used in modeling the geometry-based identification system: Identification and Train Formation.

The Identification model operated the algorithm presented in Table 1.

The Train Formation model was used to create a variety of train formations. This model used a database for a variety of railroad vehicles. The database contained the necessary geometric parameters for each vehicle extracted from associated technical documentation. The database also included all possible values of the inter-axle distance m_q^m for various railroad vehicles. Those values were sourced from technical documentation for a train fleet.

Thus, by specifying models and serial numbers of railroad vehicles, it is possible to form input data for modeling the identification process. The train speed and its changes are set by a clocking unit that determines the output speed of geometric parameters for the Identification model.

The operation of the identification system was simulated in two modes: step mode and motion mode.

In step mode, modeling takes place in free time according to clock cycles, each of which sets the beginning of an algorithm step listed in Table 1. Clocking is done manually by the operator.

Once each vehicle is identified, an output form is generated with the identification results (see an example in Fig. 4).

Comparing the identification results for individual vehicles with the input data set by the Train Formation model confirms the reliability of the proposed algorithm.

The algorithm proved reliable under conditions causing error in measuring inter-axle distances. During the modeling process, this error varied randomly in a range of ± 50 mm.

When simulating in motion mode, the Train Formation model generates a flow of geometric parameters for railroad vehicles consistent with the given train formation. These parameters enter the Identification model at a certain rate.

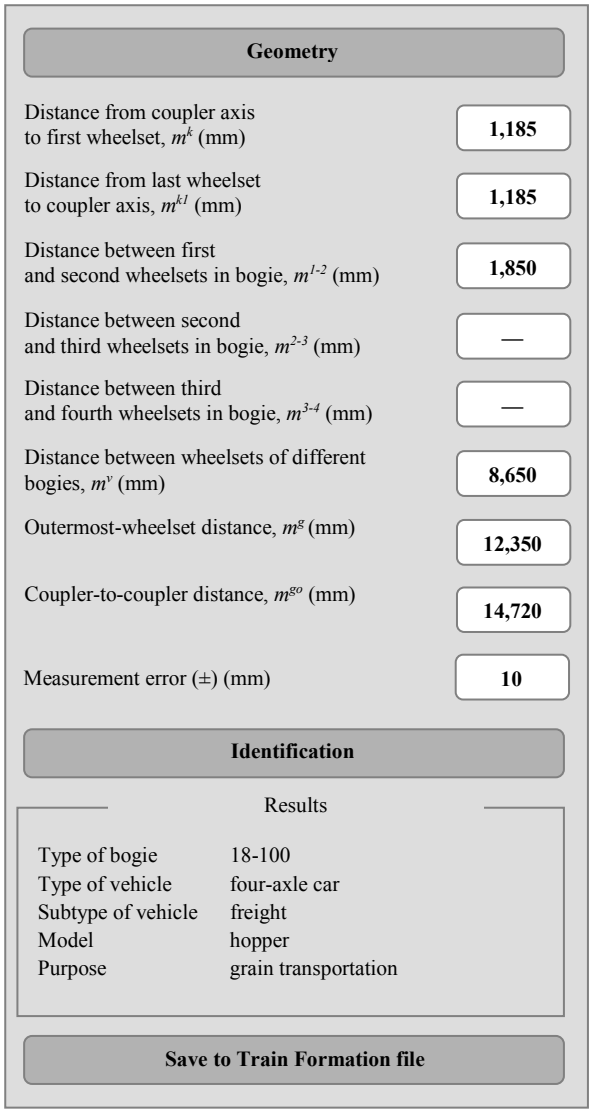


Fig. 4. An example of modeled vehicle identification in step mode.

The flow rate of geometric parameters corresponds to a specific train speed, varying in a range of 10–90 km/h. This range allows speed change patterns to be set (uniform movement, acceleration, or braking).

Once the last car passes by, the Identification model generates a Train Formation file in the form of Fig. 5.

Comparing the generated and initial Train Formation files in the models confirms the reliability of the proposed algorithm in a time scale close to real time.

Railroad vehicles					
Item no.	Type	Subtype	Model	Number of axles	Purpose
1	Locomotive	Electric	2ES5K	4	Mainline
2	Locomotive	Electric	2ES5K	4	Mainline
3	Car	Freight	Hopper	4	Grain
4	Car	Freight	Flatcar	4	Multipurpose
. .	. .	. .	. .	. .	. .
56	Car	Freight	Tank	8	Oil products
57	Car	Freight	Tank	4	Oil products

Fig. 5. Example of a Train Formation file generated in motion mode.

V. Conclusion

Paper proposed a geometry-based rolling-stock identification system that reliably identifies types of railroad vehicles when the train is traveling at a nonuniform speed. To make identification more reliable and accurate under varying speed conditions, we propose adjusting the measured overall coupler-to-coupler length.

The adjustment is based on the measured magnitude and sign of acceleration for the train and takes place automatically.

The paper described the system's algorithm, which compensates for measurement errors that occur when the train's speed changes.

Equipment based on the proposed system should be installed at intermediate stations between major junction stations. This makes the operational management of rolling stock more efficient and allows you to control the train schemes in the process of movement for compliance with applicable regulations.

Our computer-aided modeling results confirmed the efficiency of the solutions proposed.

References

[1] Gr. Theeg, and S. Vlasenko, "Railway Signalling and Interlocking. International Compendium", 3d ed., 2020, PMC Media House GmbH, Eurail press, 552 p.

[2] J. Pachl, "Railway Operation and Control", 4th ed., VTD Rail Publishing, Mountlake Terrace, 2018, 302 p.

[3] O. S. Nock, "Railway Signalling – A Treastise on the Recent Practice of British Railways", A&C Black, London, 1982, 312 p.

[4] Ansaldo STS USA Inc., MicroLok II Track Circuit PCBs, Product Catalog RSE-1D2.5, 2013.

[5] M. Ohta, "Level Crossing Obstacle Detection System Using Stereo Cameras", *Quarterly Reports of Railway Technical Research Institute*, no. 46 (2), 2005, pp. 110-117.

[6] Ye. Vesnin, V. Tsaryov, and A. Mikhaylov, "Railcar numberrecognition: solution principles and industrial applications" [in Russian], *Control Engineering Russia*, no. 1(49), 2014, pp. 60–66.

[7] N. Nishibori, T. Sasaki, and S. Hiraguri, "Development of Train System by Microwave Balises", *Quarterly Reports of Railway Technical Research Institute*, no. 43 (4), 2002, pp. 155-162.

[8] V. V. Belov, V. A. Buyanov, M. D. Rabinovich, V. F. Dudkin, B. V. Milgotin, N. M. Lyogky, and D. S. Kotletsov, "Palma: an automated vehicle identification system" [in Russian], *Zheleznodorozhny transport*, no. 8, 2002, pp. 54–59.

[9] K. Finkenzeller, "RFID Handbook: Fundamentals and Applicationin Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication", 3d ed., 2010, John Wiley & Sons. Ltd, 462 p.

[10] D. Bandara, T. Melaragno, D. Wijesekara, and P. Costa, "Multi-Tiered Cognitive Radio Network for Positive Train Control Operations", *Proceedings Joint Rail Conference (JRC)*, 2016, ASME, Columbia, pp. 1-10.

[11] M. V. Romkin, "Rolling-stock identification device" [in Russian],*Utility model patent*, no. 78159 of November 20, 2008.

[12] A. A. Mironov. "Monitoring and diagnostic tools" [in Russian]. *Avtomatika, svyaz, informatika*, no. 12, 2012, pp. 44–46.

[13] V. A. Zasov, and M. V. Romkin, "Rolling-stock identification device"[in Russian], *Utility model patent*, no. 154205 of October 6, 2014.

[14] A. Cichocki, and Sh. Amari, "Adaptive Blind Signal and Image Processing: Learning algorithms and applications, John Wiley & Sons, Ltd, 2002, 555 p.

[15] P. S. Deniz, "Adaptive filtering algorithms and practical implementation", 3d ed., Springer Science + Business Media, New York, 2008, 627 p.

[16] F. Pointner, and H. Kalteis, "Reliable wheel sensors as the basis for highly available systems", *Signal+Draht*, no. 4, 2017, pp. 10-16.

Markov Model of Quantized Speech Signal

Prozorov D.E.
Vyatka State University
Kirov, Russia
de_prozorov@vyatsu.ru

Metelyov A.P.
Vyatka State University
Kirov, Russia
ap_metelev@vyatsu.ru

***Abstract*—Currently, automatic speech recognition systems are widely used. When developing such systems, significant attention is paid to choice of parameterization method of speech signals. The paper presents options for representing fragments of speech signals by simple and high-order Markov chains. Special cases of the implementation of these models are also presented. The results show that when using the methods of Markov parameterization in problems of automatic recognition of speech commands, it is advisable to use models of high order Markov chains. The considered models of speech signals can significantly reduce the requirements on the computing resources of automatic speech recognition systems.**

Keywords—speech recognition, Markov chain, parameterization method, speech signals, transition probability matrix.

I. INTRODUCTION

Significant attention is paid to the choice of the parameterization method of speech signals in the development of systems for automatic recognition of speech commands. A large number of parametrization models and methods of speech signals have been developed that provide high quality systems for automatic recognition of speech commands, among which the most common methods are based on the calculation of cepstral coefficients and linear speech prediction coefficients. [1-4].

However, it cannot be said that the parameterization problem is completely solved. The performance of automatic speech recognition systems is still far from the "performance" of the human auditory system. For example, in the problem of numbers recognition, when the dictionary is small and a significant part of the resources is spent on acoustic modeling, the performance of automatic speech recognition systems is much lower than the performance of a human [2]. This is partly due to the large number of hidden and explicit state variables of modern systems for automatic recognition of speech commands. Thus, the task of developing models and methods of parameterization of speech signals, allowing to find a compromise between performance, resource requirements and quality of work of systems for automatic recognition of speech commands is actual.

The first works devoted to Markov models of speech signals appeared in the 1970s. So, in the monograph [5], a detailed analysis of Markov models of delta-modulated speech signals was performed. Further studies showed some disadvantages of linear delta modulation, therefore more efficient coding methods and presentation formats of speech signals were developed [6]. Nevertheless, a small number of publications devoted to the study of Markov parameterization methods for speech signals and such advantages of Markov models as a relatively small number of model parameters and the linear computational complexity of parameter calculation algorithms continue to arouse interest in these methods.

In this paper, Markov models of quantized speech signals are considered.

II. FORMULATION OF THE PROBLEM

Let a fragment of a speech signal be represented by an array of discrete samples

$$\mathbf{s} = \left\{ s^{(1)} s^{(2)} \ldots s^{(L)} \right\}, \tag{1}$$

where $s^{(k)}$ – k-th binary sample of the form

$$s^{(k)} = \sum_{b=0}^{N-1} \left({}^{b}s^{(k)} \cdot 2^{N-b-1} \right), \; k = \overline{1,L}, \; {}^{b}s^{(k)} = \overline{0,1}. \tag{2}$$

L – fragment length.

Fragment (1) can be considered as a superposition of bit sequences ${}^{b}\mathbf{s}$

$$\mathbf{s} = \sum_{b=0}^{N-1} \left({}^{b}\mathbf{s} \cdot 2^{N-b-1} \right). \tag{3}$$

where $\{\cdot\}$ – scalar multiplication operation, ${}^{b}\mathbf{s}$ – binary sequence formed by the b-th bit of binary samples of a fragment of a speech signal.

A series of experiments [5,7,8] shows that for speech signals the assumption is valid that the correlation between the samples $s^{(n)}$ and $s^{(n-k)}$ of the speech signal decreases substantially with an increase of the interval k between the samples. Therefore, it is possible to indicate an interval beyond which correlation relationships practically do not exist. Given this assumption, the mathematical apparatus of simple or complex Markov chains can be used to describe and analyze short fragments of speech signals [9, 10].

It is required to develop a model and method of parameterization of speech signals fragments (1).

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

III. HIGH ORDER MARKOV CHAIN

Consider a high order (connected with m) Markov chain with $l = 2^N$ states and transition probabilities of the form (4)

$$P\left(s_n^{(k)} \mid s_i^{(k-1)} s_j^{(k-2)} \ldots s_r^{(k-m)} s_q^{(k-m-1)}, \ldots\right) = \\ = P\left(s_n^{(k)} \mid s_i^{(k-1)} s_j^{(k-2)} \ldots s_r^{(k-m)}\right), \quad (4)$$

where $i, j, r, q, n = \overline{1, l}$, $k = \overline{1, L}$.

We pass from the high order Markov chain to the simple one, forming a vector $\vec{s}^{(k-1)} = \left(s_1^{(k-1)} s_2^{(k-2)} \ldots s_n^{(k-m)}\right)$ of length *m* [11].

Then

$$P\left(\vec{s}^{(k)} \mid \vec{s}^{(k-1)} \ldots \vec{s}^{(k-m)}\right) = \\ = P\left(s_n^{(k)}, s_i^{(k-1)}, \ldots, s_v^{(k-m+1)} \mid s_i^{(k-1)}, s_j^{(k-2)}, \ldots, s_r^{(k-m)}\right), \quad (5)$$

where $i, j, n, r, v = \overline{1, l}$.

The number of states of the simple Markov chain transformed in this way is l^m.

With known conditional probabilities (4), we can determine the transition probabilities

$$P\left(\vec{s}^{(k)} \mid \vec{s}^{(k-1)}\right) = \\ = \frac{P\left(\left\{s_n^{(k)}, \ldots, s_v^{(k-m+1)}\right\}, \left\{s_i^{(k-1)}, \ldots, s_r^{(k-m)}\right\}\right)}{P\left(s_i^{(k-1)}, \ldots, s_r^{(k-m)}\right)} = \\ = \frac{P\left(s_n^{(k)} s_i^{(k-1)}, \ldots, s_r^{(k-m)}\right)}{P\left(s_i^{(k-1)}, \ldots, s_r^{(k-m)}\right)} = P\left(s_n^{(k)} \mid s_i^{(k-1)}, \ldots, s_r^{(k-m)}\right), \quad (6)$$

where $i, j, n, r, v = \overline{1, l}$.

For example, for a Markov chain of connection $m = 2$ vector states $\vec{s}^{(k)}$ take values from the set $\left\{\left\{s_1, s_1\right\}, \left\{s_1, s_2\right\}, \left\{s_2, s_1\right\}, \left\{s_2, s_2\right\}\right\}$ and the transition probability matrix has the form:

$$\Pi = \begin{array}{c} \\ s_1 s_1 \\ s_1 s_2 \\ s_2 s_1 \\ s_2 s_2 \end{array} \begin{array}{c} \begin{array}{cccc} \underline{s_1 s_1} & \underline{s_1 s_2} & \underline{s_2 s_1} & \underline{s_2 s_2} \end{array} \\ \begin{bmatrix} \pi_{111} & \pi_{112} & 0 & 0 \\ 0 & 0 & \pi_{121} & \pi_{122} \\ \pi_{211} & \pi_{212} & 0 & 0 \\ 0 & 0 & \pi_{221} & \pi_{222} \end{bmatrix} \end{array}. \quad (7)$$

where $\pi_{ijn} = P\left(s_n^{(k)}, s_i^{(k-1)} \mid s_i^{(k-1)}, s_j^{(k-2)}\right)$, $i, j, n = \overline{1, 2}$.

Zero elements of the transition probability matrix correspond to the probabilities of impossible events.

In a similar way, we can obtain transition probability matrices for the general case ($m > 2$, $l > 2$). For matrix elements of the form (7), modified normalization conditions must be observed

$$\sum_{n=1}^{l} \pi_{rv\ldots in} = 1, \quad i, j = \overline{1, l}, \quad (8)$$

and coherence

$$p_{v\ldots in} = \sum_{r=1}^{l} p_{rv\ldots i} \pi_{rv\ldots in}, \quad v, \ldots, i, n = \overline{1, l}, \quad (9)$$

where $p_{v\ldots in} = P\left(s_n^{(k)}, s_i^{(k-1)}, \ldots, s_v^{(k-m+1)}\right)$ – unconditional probability of a combination of *m* states $\left\{s_v \ldots s_i s_n\right\}$.

The difference equations are valid for the simple vector Markov chain introduced in this way.

$$\mathbf{p}^{(k)} = \mathbf{p}^{(k-1)} \Pi = \mathbf{p}^{(0)} \Pi^k, \quad (10)$$

where $\mathbf{p}^{(k)}$ – vector of unconditional probabilities of all possible combinations of *m* states at the k-th step.

Consider the methods of Markov parameterization of speech signals fragments using the described model in order of increasing complexity.

IV. PARAMETRIZATION BY A SIMPLE MARKOV CHAIN

In the general case, the method of parametrization of a speech signal fragment by a simple Markov chain contains the following steps sequence.

1. The bit sequences ${}^{i}\mathbf{s}$ ($i = \overline{1, N}$) are combined into disjoint arrays $\left[{}^{b}\mathbf{s}, {}^{b+1}\mathbf{s}, \ldots, {}^{b+N_b-1}\mathbf{s}\right]^{\mathrm{T}}$ of size $N_b \times L$.

Sequences of meta states ${}^{b}S^{(k)}$ are considered as simple Markov chains ${}^{b}\mathbf{S}$, $b = \overline{0, N/N_b - 1}$ with the number of states equal to 2^{N_b}.

$$\left\{\begin{pmatrix} {}^{b}S^{(1)} \\ \vdots \\ {}^{b+N_b-1}S^{(1)} \end{pmatrix}, \begin{pmatrix} {}^{b}S^{(2)} \\ \vdots \\ {}^{b+N_b-1}S^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} {}^{b}S^{(L)} \\ \vdots \\ {}^{b+N_b-1}S^{(L)} \end{pmatrix}\right\} = \quad (11) \\ = \left\{{}^{b}S^{(1)}, {}^{b}S^{(2)}, \ldots, {}^{b}S^{(L)}\right\} = {}^{b}\mathbf{S}$$

2. Each vector ${}^{b}\mathbf{S}$ is divided into p intersecting segments ${}^{b}_{1}\mathbf{S}, \ldots, {}^{b}_{p}\mathbf{S}$ with overlap coefficient *h*.

3. The elements ${}^{b}_{q}\pi_{ij}$ of the transition probability matrices ${}^{b}_{q}\Pi$ of a simple Markov chain are estimated for each *q*-th segment of the vector ${}^{b}\mathbf{S}$.

$${}^{b}_{q}\pi_{ij} = \frac{\hat{P}\left({}^{b}_{q}s_j^{(k)}, {}^{b}_{q}s_i^{(k-1)}\right)}{\hat{P}\left({}^{b}_{q}s_i\right)} = \hat{P}\left({}^{b}_{q}s_j^{(k)} \mid {}^{b}_{q}s_i^{(k-1)}\right), \quad (12)$$

where $b = \overline{0, N/N_b - 1}$, $q = \overline{1, p}$, $k = \overline{1, L}$, $i, j = \overline{1, 2^{N_b}}$; $\hat{P}\left({}^{b}_{q}s_j^{(k)}, {}^{b}_{q}s_i^{(k-1)}\right)$ – relative frequency of combinations $\left\{{}^{b}_{q}s_j^{(k)}, {}^{b}_{q}s_i^{(k-1)}\right\}$ in the *q*-th segment of the vector ${}^{b}\mathbf{S}$; $\hat{P}\left({}^{b}_{q}s_i\right)$ –

relative frequency of the i-th state of the Markov chain in the q-th segment of the vector ${}^{b}\mathbf{S}$.

4. The resulting transition probability matrices ${}^{b}_{q}\Pi$ are combined into a signal parameters matrix.

$$\mathbf{M} = \begin{bmatrix} {}^{0}_{1}\Pi & {}^{0}_{2}\Pi & \dots & {}^{0}_{p}\Pi \\ \dots & \dots & \dots & \dots \\ {}^{N/N_b-1}_{1}\Pi & {}^{N/N_b-1}_{2}\Pi & \dots & {}^{N/N_b-1}_{2}\Pi \end{bmatrix}. \qquad (13)$$

The diagram of possible transitions of a simple Markov chain ${}^{b}\mathbf{S}$ for $N_b = 2$ is shown in Fig. 1.

Example. $m = 2$, $N = 4$, $N_b = 2$.

1. Binary bit sequences ${}^{0}\mathbf{s},\dots,{}^{3}\mathbf{s}$ are combined in sequences ${}^{b}\mathbf{S}$ (Fig. 2) such that

$${}^{0}\mathbf{S} = \left\{\begin{pmatrix} {}^{0}s^{(1)} \\ {}^{1}s^{(1)} \end{pmatrix}, \begin{pmatrix} {}^{0}s^{(2)} \\ {}^{1}s^{(2)} \end{pmatrix}, \dots, \begin{pmatrix} {}^{0}s^{(L)} \\ {}^{1}s^{(L)} \end{pmatrix}\right\} = \qquad (14)$$
$$= \left\{{}^{0}\mathbf{S}^{(1)}, {}^{0}\mathbf{S}^{(2)}, \dots, {}^{0}\mathbf{S}^{(L)}\right\},$$

$${}^{1}\mathbf{S} = \left\{\begin{pmatrix} {}^{2}s^{(1)} \\ {}^{3}s^{(1)} \end{pmatrix}, \begin{pmatrix} {}^{2}s^{(2)} \\ {}^{3}s^{(2)} \end{pmatrix}, \dots, \begin{pmatrix} {}^{2}s^{(L)} \\ {}^{3}s^{(L)} \end{pmatrix}\right\} =$$
$$= \left\{{}^{1}\mathbf{S}^{(1)}, {}^{1}\mathbf{S}^{(2)}, \dots, {}^{1}\mathbf{S}^{(L)}\right\}.$$

The number of meta states ${}^{b}\mathbf{S}^{(k)}$ of such sequences is $2^{N_b} = 4$.

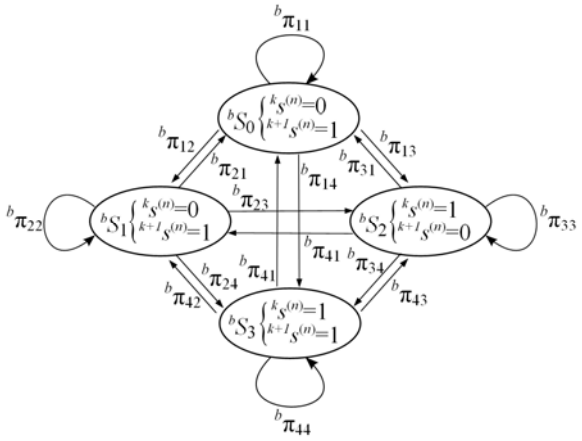


Fig. 1. The transition diagram of the Markov chain ${}^{b}\mathbf{S}$

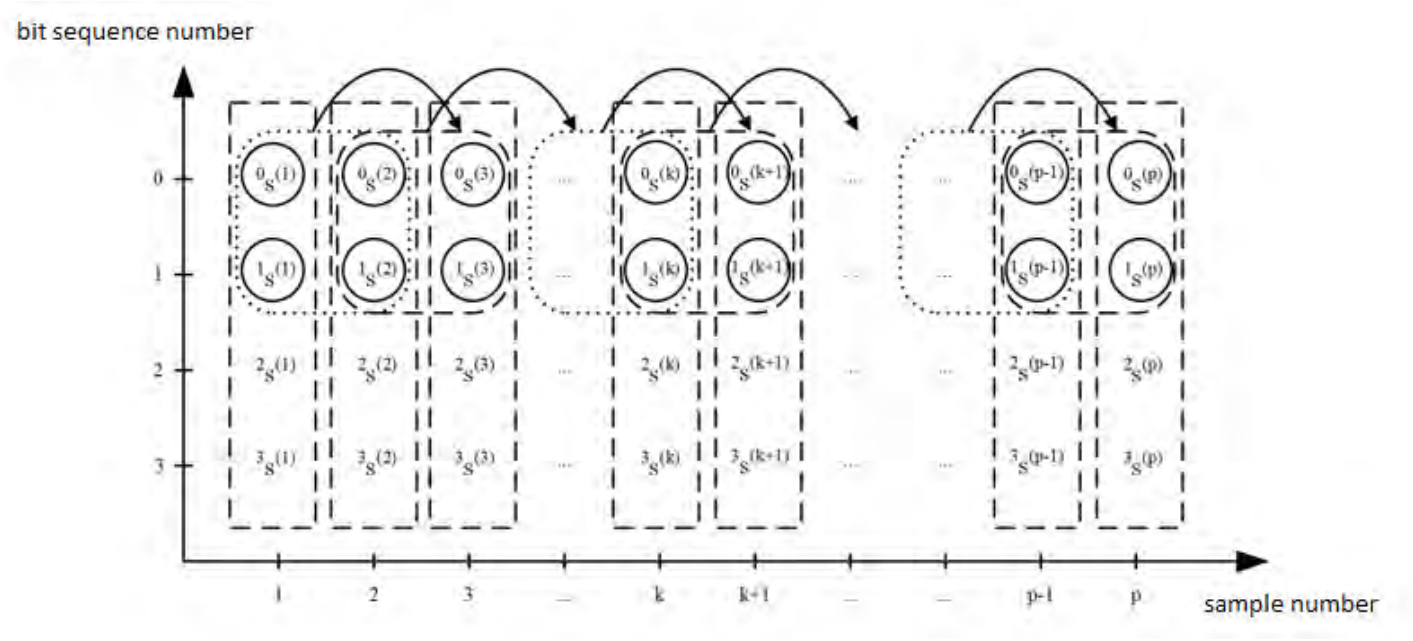


Fig. 2 Plots of transitions of the multi-valued Markov chain for $m = 2$ and $N_b = 2$.

Assuming that the sequences ${}^{b}\mathbf{S}$ are high order Markov chains of connectedness $m = 2$ and the number of states is four

$${}^{b}S_1 = \begin{pmatrix} {}^{b}s_1 \\ {}^{b}s_1 \end{pmatrix}, \; {}^{b}S_2 = \begin{pmatrix} {}^{b}s_1 \\ {}^{b}s_2 \end{pmatrix}, \; {}^{b}S_3 = \begin{pmatrix} {}^{b}s_2 \\ {}^{b}s_1 \end{pmatrix}, \; {}^{b}S_4 = \begin{pmatrix} {}^{b}s_2 \\ {}^{b}s_2 \end{pmatrix}, \qquad (15)$$

we transform them into simple Markov chains with the number of states equal to $2^{mN_b} = 16$, forming vectors

$${}^{b}\vec{S}^{(k)} = \left({}^{b}S^{(k)}, {}^{b}S^{(k-1)}\right) \qquad (16)$$

and vector sequences ${}^{b}\vec{\mathbf{S}}$ (14).

2. Sequences ${}^{b}\vec{\mathbf{S}}$ are divided into p intersecting segments with overlap coefficient h: ${}_{1}^{1}\vec{\mathbf{S}},\ldots,{}_{p}^{1}\vec{\mathbf{S}}$ и ${}_{1}^{2}\vec{\mathbf{S}},\ldots,{}_{p}^{2}\vec{\mathbf{S}}$.

3. For each q-th segment of the sequence ${}^{b}\vec{\mathbf{S}}$ elements ${}_{q}^{b}\pi_{ijn}$ of transition probability matrices ${}_{q}^{b}\Pi$ are estimated

$$ {}_{q}^{b}\pi_{ijn} = \frac{P\left({}_{q}^{b}\vec{S}^{(k)}, {}_{q}^{b}\vec{S}^{(k-1)}\right)}{P\left({}_{q}^{b}\vec{S}^{(k-1)}\right)} = P\left({}_{q}^{b}S_{n}^{(k)} \mid {}_{q}^{b}S_{j}^{(k-1)}, {}_{q}^{b}S_{i}^{(k-2)}\right) = $$

$$ = \frac{\hat{P}\left({}_{q}^{b}S_{n}^{(k)}, {}_{q}^{b}S_{j}^{(k-1)}, {}_{q}^{b}S_{i}^{(k-2)}\right)}{\hat{P}\left({}_{q}^{b}S_{j}^{(k-1)}, {}_{q}^{b}S_{i}^{(k-2)}\right)}, \qquad (17) $$

where $b = \overline{0,1}$, $q = \overline{1,p}$, $k = \overline{1,L}$, $i,j,n = \overline{1,4}$.

4. Nonzero elements ${}_{q}^{b}\pi_{ijn}$ of the transition probability matrix ${}_{q}^{b}\Pi$ are combined into a signal parameter matrix $\mathbf{M}$.

V. Experiment

The speech signal models considered in the paper are used to evaluate the effectiveness of Markov parameterization methods in solving the problem of speaker-dependent recognition of speech commands. As an alternative method of speech signals parameterization, the method of parameterization with cepstral coefficients is used (MFCC).

For the experiment, the author's collection of 150 speech commands with a sampling frequency of 8 kHz was formed.

The experiment consisted of the following steps:

1) estimation of the average parameterization time of one fragment of a speech signal;

2) probability estimation of correct recognition of commands as a result of performing 1000 experiments for each model (Table 1).

The experimental results are presented in table 1. The following abbreviations are introduced to indicate the model used for parameterization: common model name - dtmc (Discrete Time Markov Chain), the next two digits indicate the order of the Markov chain - m and the number of grouped bit sequences - N_b. The numbers of the most significant bits used for parameterization are indicated in parentheses.

The experiment showed that a compromise between the recognition efficiency of speech commands and the computational complexity of the parameterization algorithm is achieved by using (to estimate the parameters of the approximating Markov chain) from one to four high bits of a linearly quantized speech signal.

The results indicate a significant (16-19 times) reduction in time spent on parameterization of fragments of speech signals with a concomitant decrease in the probability of commands recognition by 5-8% relative to the classical method of parameterization with cepstral coefficients.

TABLE I. Probability of recognition and time of parametrization

	Parameterization method	Relative parameterization time for one fragment	Recognition probability,%
Simple chain	dtmc11(0)	0,052	76,3
	dtmc11(0,1)	0,055	83,5
	dtmc11(0,1,2)	0,058	84,2
	dtmc11(0,1,2,3)	0,062	84,7
	dtmc12(0,1)	0,051	85,3
	dtmc12(0,1,2,3)	0,055	89,7
High order chain	dtmc21(0)	0,052	83,4
	dtmc21(0,1)	0,055	92,4
	dtmc21(0,1,2)	0,059	94,8
	dtmc21(0,1,2,3)	0,062	95,8
	dtmc22(0,1)	0,052	85,6
	dtmc21(0,1,2,3)	0,055	93,9
	MFCC	1	99,7

VI. Conclusion

The paper summarizes the options for representing fragments of speech signals by simple and high order Markov chains. Particular cases of the implementation of these models for the parameterization of speech signals by Markov chains with $m = 1$ and $m = 2$ are presented.

It is shown that when using the methods of Markov parameterization in problems of automatic speaker-dependent recognition of speech commands, it is advisable to use models of high order Markov chains. The considered models of speech signals can significantly reduce the requirements on the computing resources of automatic speech recognition systems.

References

[1] A.V. Oppenheim, R.W. Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1975.

[2] Picone, J.W., "Signal modeling techniques in speech recognition", proceedings of the IEEE, September 1993, pp. 1215-1247.

[3] X. Huang, A.Acero, H.Hon., Spoken Language Processing: A guide to theory, algorithm, and system development, Prentice Hall, 2001.

[4] L.R. Rabiner, Digital processing of speech signal, Moscow, 1981.

[5] Venediktov M.D., Zhenevskiy YU.P., Markov V.V., Delta Modulation. Theory and Application, Moscow, 1976, pp. 104-114.

[6] S. G. Richter, Voice coding and transmission in digital mobile radio systems, Moscow, 2010.

[7] Pletnev K.V., Prozorov D.E., "Analysis of the Markov parameterization method of speech signals", Information Systems and Technologies, 2014, №1(81), pp. 24-29.

[8] Pletnev K.V., "Method of parameterization of speech signals by second-order Markov chains", proceedings «Theory and practice of modern science», Moscow, 2013, vol. 1, pp. 193-198.

[9] J. Kemeny, J. Snell, Finite Markov chains, Moscow, 1970.

[10] Yanshin V.V., "High order Markov chains and their properties", Radio engineering and electronics, 1993, vol. 38, № 6, pp. 1081-1091.

Investigation of a Broadband Five-Stub 3 dB Coupler Using Microstrip Cells

Denis A. Letavin
Ural Federal University
Yekaterinburg, Russia
d.a.letavin@urfu.ru

Ilya A. Terebov
Ural Federal University
Yekaterinburg, Russia
ilyaterebov@yandex.ru

***Abstract*—Diagram-forming circuits contain passive microwave devices, such as phase shifters, directional couplers. Various requirements are imposed on such devices, for example, low cost, low weight and size, etc. The layout of a broadband 3 dB coupler functioning at a frequency of 1 GHz is studied. The coupler miniaturization was realized by exchanging the quarter-wave segments with microstrip cells having comparable characteristics with smaller dimensions. Such a replacement, allowed reducing the space of the coupler by more than 3 times, relative to the standard design.**

Keywords—filter, coupler, miniaturization, device.

I. INTRODUCTION

Depending on the conditions of use of microwave devices, certain parameters are gaining importance: dimensions, permissible operating temperatures, bandwidth, etc. In this work, a compact directional coupler is considered. A coupler is a device used to divide or sum power. The dimensions of such devices are linearly related to the operating frequency, the lower it is, the greater the area they occupy on a microwave substrate. This is especially noticeable when the device is operating in the decimeter range of frequencies. In this regard, it is worthwhile to find an approach that allows not only to decrease the sizes of the scheme, but also to maintain its operability at the level of the usual design. Today, in the IEEE database, you can find a large number of works devoted to the miniaturization of various microwave devices. Compact couplers with different implementations are described in [1] - [14]. In [1] the author proposes a design of a compact divider, assembled on lumped elements installed in place of transmission lines. In [2,3], a mathematical approach to the construction of such structures is shown using an example of a circular directional coupler. Work [4] demonstrated the possibility of using a slit between two transmission lines located on different substrate layers to miniaturize a directional coupler. In [5,8], P-circuits are used in which the extreme capacities are combined with the capacities of neighboring circuits. Work [6] describes the design of a compact ring coupler with an original way to optimize the dimensions of the device. The work [7] shows a design implemented on T-circuits, which can be simply manufactured using standard methods for manufacturing printed circuit boards. In [9], bends of transmission lines and vias installed between them are used. In [10], artificial transmission lines of original design are used to miniaturize the power divider. In [11], the authors use low-pass filters to miniaturize a circular directional coupler. Each method allows to achieve different miniaturization efficiency. In our case, microstrip cells will be used, which have comparable characteristics at the central frequency and its surroundings with the characteristics of planar segments.

II. DESIGN COMPACT COUPLER

The target of this work is to obtain a compact directional coupler with a wide bandwidth and well-realized dimensions. To achieve this goal, the following tasks are solved: designing a standard coupler to obtain its dimensions and characteristics, with which a comparison will be made later; calculation and synthesis of microstrip cells with comparable characteristics with replaceable segments; modeling a compact coupler and comparing the resulting characteristics with the characteristics of a conventional coupler model; making a model of the device under development and measuring its characteristics.

To obtain a wide frequency band, additional stubs are installed to the usual branch-line coupler. However, the number of connected loops is limited in that their wave impedance increases and it becomes difficult to put them into practice. The five-stub 3 dB coupler has in its design loops with a wave impedance of 100 Ohms (Fig. 1). This coupler provides an equal separation of power separation the outputs with a phase difference of 90 degrees. For this reason, the device is symmetrical, then this principle of operation will be performed when microwave power is applied to any of the coupler inputs. Initially, according to the scheme shown in Fig. 1 in the Cadence AWR DE 14 program, the coupler topology on standard microstrip segments and FR4 substrate was assembled. The obtained model of the coupler with an frequency of 1 GHz is shown in Fig. 2, and its characteristics after calculation are shown in Fig. 3, 4. The area of the coupler is 2757 mm^2.

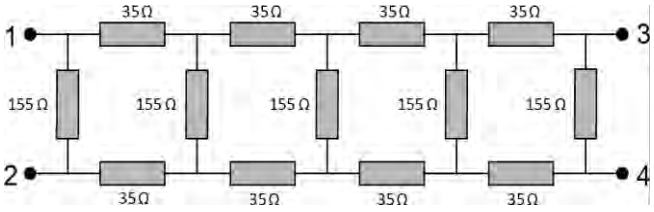


Fig. 1. The sheme of the coupler

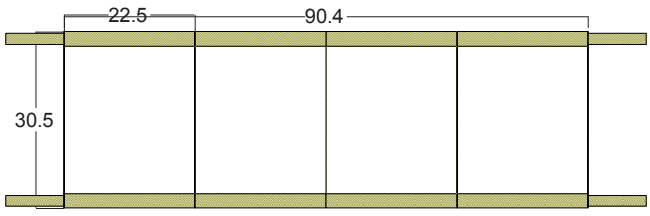


Fig. 2. AWR standard branch layout

Figure 2 shows that a conventional coupler has an area inside the figures described by quarter-wave segments, which is not used in any way. The simple bending of the λ/4 segments does not essential decrease the scope of the device; therefore, it is necessary to replace standard segments with compact structures, while it is necessary to avoid a significant modification in the characteristics of the device.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

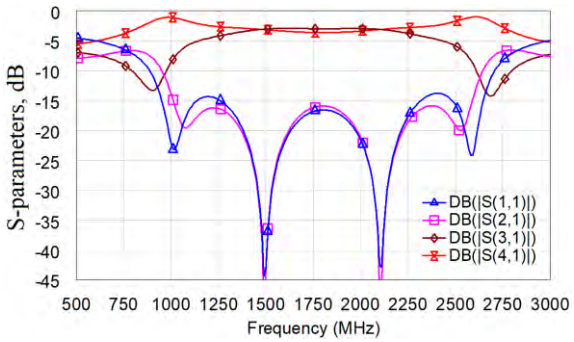


Fig. 3. Frequency dependence of S-parameters for a standard coupler obtained in AWR

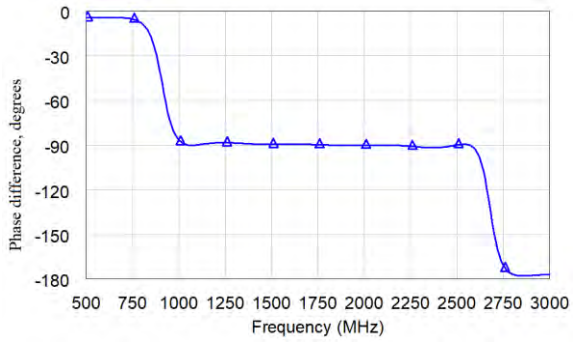


Fig. 4. Graph showing the change in the phase difference between the output signals of the splitter received by AWR

The standard coupler works atin the frequency band (by isolation level -15 dB) 1000 - 2580 MHz. In the same frequency band, a phase difference of 90 degrees is made between the output signals. At the same time, the shoulder balance of +0.5 dB is preserved only in the frequency band 1400-2200 MHz. After the design of the coupler was obtained, a miniaturization of the device was performed, namely, replacing the quarter-wave segments with an analog in the form of microstrip cells, pre-calculated and optimized for the required requirements. To grow the miniaturization efficiency, the internal space of the coupler, the wide elements of the cells were located in the internal area of the coupler. The layout of such a coupler is shown in Fig. 5. The size of the device with this arrangement turned out to be equal to 833 mm2, which is more than 3 times less than the usual implementation of the branch (69.8%). The graph of S-parameters and phase versus frequencies are obtained after numerical analysis in the AWR program (Fig. 6,7).

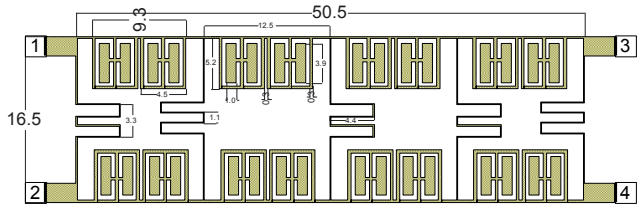


Fig. 5. AWR compact coupler layout

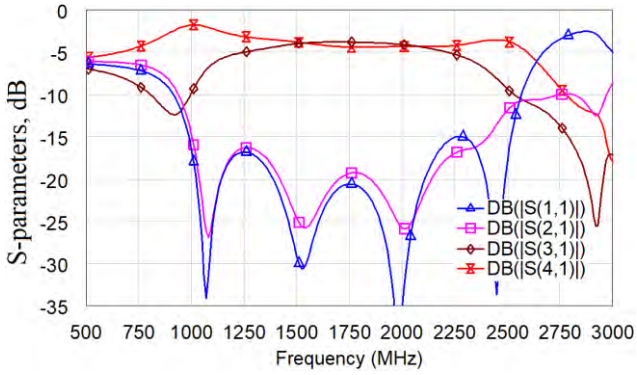


Fig. 6. Frequency-dependent S-parameters for a compact coupler obtained in AWR

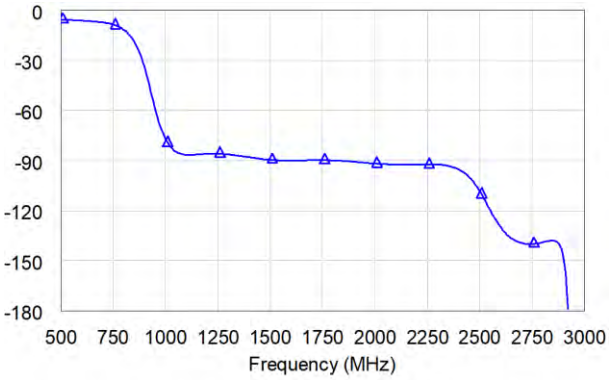


Fig. 7. A graph of the phase difference between the output signals of the compact coupler obtained by AWR

The compact coupler works in the frequency band (estimated by isolation level -15 dB) 1000 - 2420 MHz. In this case, a phase difference of 90 degrees is performed in a smaller frequency band. In this case, the shoulder balance of +0.5 dB is preserved only in the frequency band 1450-2180 MHz.

III. Prototype Measurements

To verify the correctness of the calculations, a coupler model was made using photolithography; its photograph is shown in Fig. 8. To measure the S-parameters, the Rode&Shwarz ZVA24 vector network analyzer was used. The removed experimental dependences are shown in Figs 9.10.

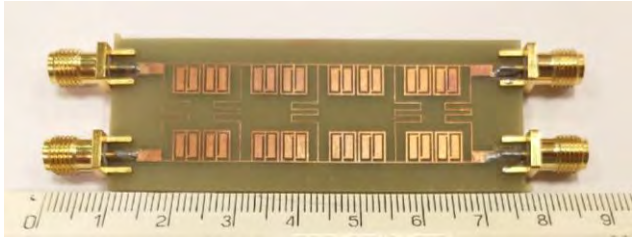


Fig. 8. The layout of the compact coupler

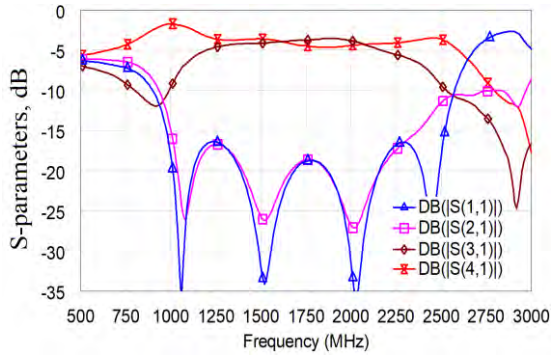


Fig. 8. Frequency-dependent measured S-parameters for a compact coupler

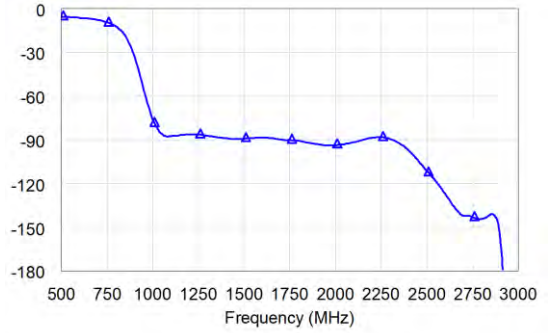


Fig. 9. Graph showing the change in the phase difference between the output signals of the splitter received by AWR

The layout works in the frequency band (estimated by isolation level -15 dB) 1000 - 2380 MHz. In this case, the phase difference of 90+3 degrees is performed in the band 1050 - 2325 MHz. The imbalance between gear ratios has increased. Deterioration of characteristics in comparison with the calculated data can be caused by manufacturing errors and the presence of connectors that were not taken into account in the simulation. For a simpler comparison, the main characteristics of the devices are presented in Table 1.

TABLE I. COMPARISON OF STANDARD AND MINIATURE COUPLERS

Parameters	*Standard*	*Compact*
bandwidth, MHz	1580	1380
Relative bandwidth, %	88.3	81.6
area, mm^2	3786	833
Relative area, %	100	30.2
Central frequency, MHz	1790	1690
The phase outputs,°	90	89.5

IV. CONCLUSION

An implementation option for a compact broadband coupler has been obtained. Miniaturization was conceded out based on the use of cells, whose characteristics are comparable to the characteristics of microstrip segments, but at the same time take up less space on the substrate. The simulation was performed in the Cadence AWR DE 14 program, and the model was made using photolithography. The proposed coupler has a simply implemented topology, and its area is more than 3 times smaller than the area of a conventional coupler configured on the same frequency. Negative factors of miniaturization are an increase in losses in the passband, growth in the imbalance between transmission coefficients, and a decrease in the bandwidth.

ACKNOWLEDGMENT

The work was supported by Act 211 Government of the Russian Federation, contract № 02.A03.21.0006.

REFERENCES

[1] Bayaner Arigong, Mi Zhou, Han Ren, Chang Chen and Hualiang Zhang, “A Compact Lumped-Component Coupler with Tunable Coupling Ratios and Reconfigurable Responses,” 2018 IEEE/MTT-S International Microwave Symposium – IMS. DOI: 10.1109/MWSYM.2018.8439297

[2] Slawomir Koziel, Ari and Ari T. Sigurðsson and Freysteinn V. Vidarsson, “Accurate Design-Oriented Modeling of Compact Microwave Couplers in Constrained Domains,” 2018 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO). DOI: 10.1109/NEMO.2018.8503172.

[3] Slawomir Koziel, Adrian Bekasiewicz and John Bandler, “Rapid dimension scaling of compact microwave couplers with power split correction,” 2017 47th European Microwave Conference (EuMC). DOI: 10.23919/EuMC.2017.8230865

[4] S.F. Ausordin, S.K.A. Rahim, Norhudah Seman and R. Dewan, “A compact 3-dB coupler on a dual substrate layer with a rectangular slotted microstrip ground plane,” 2013 IEEE Business Engineering and Industrial Applications Colloquium (BEIAC). DOI: 10.1109/BEIAC.2013.6560103.

[5] J. Sung-Chan, et al., "A Design Methodology for Miniaturized 3-dB Branch-Line Hybrid Couplers Using Distributed Capacitors Printed in the Inner Area," Microwave Theory and Techniques, IEEE Transactions on, vol. 56, pp. 2950-2953, 2008.

[6] Slawomir Koziel and Adrian Bekasiewicz, “Novel structure and size-reduction-oriented design of microstrip compact rat-race coupler,” 2016 IEEE/ACES International Conference on Wireless Information Technology and Systems (ICWITS) and Applied Computational Electromagnetics (ACES). DOI: 10.1109/ROPACES.2016.7465390.

[7] S.-S. Liao, P.-T. Sun, N.-C. Chin, and J.-T. Peng, “A novel compact-size branch-line coupler,” IEEE Microw. Wireless Comp. Lett., vol. 15, no. 9, pp. 588-590, Sep. 2005.

[8] K.-Y. Tsai, H.-S. Yang, J.-H. Chen, and Y.-J.Chen, “A Miniaturized 3 dB Branch-Line Hybrid Coupler with Harmonics Suppression,” IEEE Microw. Wireless Comp. Lett., vol. 21, no. 10, pp. 537–539, Oct. 2011.

[9] Vidhish Vedala, Abhishek Shandiliya and Madhur Deo Upadhayay, “Compact Horse Shoe Shape Rat-Race Coupler,” 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN). DOI: 10.1109/SPIN48934.2020.9071351.

[10] C. H. Tseng and H. J. Chen, “Compact rat-race coupler using shuntstub based artificial transmission lines,” IEEE Microw. Wireless Compon. Lett.,vol. 18, no. 11, pp. 734–736,Sep.2008.

[11] Letavin, D.A., Mitelman, Y.E. and Chechetkin, V.A., “A Novel Simple Miniaturization Technique for Microstrip Couplers,” 8th International Conference on Mathematical Modeling in Physical Science, IC-MSQUARE 2019. DOI: 10.1088/1742-6596/1391/1/012110.

Appraisal of the Effective Number of Bits of the ADC for Sensors with Account for Dynamic Errors

Leonty Samoilov
Southern Federal University,
Rostov-on-Don, Russia
lksamoilov@sfedu.ru

Darya Denisenko
Don State Technical University,
Southern Federal University,
Rostov-on-Don, Russia
d.y.denisenko@gmail.com

Nikolay Prokopenko
Member, IEEE
Don State Technical University,
Institute for Design Problems in Microelectronics of RAS,
Rostov-on-Don, Zelenograd, Russia
prokopenko@sssu.ru

***Abstract*— The static maximum relative error of the ADC for sensors is determined by the value of its least significant bit. During the input signal conversion in the ADC the information delay causes the dynamic error, which is added to the static error and actually reduces the effective number of bits. This decrease can be estimated by the effective number of bits, which illustrates the qualitative and quantitative assessment of the additional error. The article demonstrates that the effective bits rate is defined by the delay period of information in the ADC, as well as the sampling rate of the input signal. This, in turn, requires setting the spectrographic density (SD) for the converted signal, as well as an algorithm for selecting the sampling rate. It is recommended that the delay time of the ADC includes the delay in the analog storage and amplitude multiplexer, which are located at the ADC input and are directly involved in the conversion process. The article draws attention to the fact that the real sampling rate of analog signals, chosen from the point of view of the straight task of errors distribution in the automatic control system, significantly exceeds the sampling rate according to the sampling theorem. This requires the correction of the obtained results using the sampling theorem, in the direction of reduction of effective number of bits.**

Keywords—automatic control system, dynamic error, information delay, flash ADC, pipelined ADC, serial ADC, sensors, signal spectrum

I. Introduction

In the control and monitor systems (CMS), the ADC for sensors is one of the main units that determine the error and system performance.

By the structure, the ADCs can be divided into the flash, serial and parallel-serial (pipelined) ones.

Static maximum reduced error of the q – bit ADC (γ_{st}) is equal to the least significant bit:

$$\gamma_{st} = \frac{1}{2^q}. \qquad (1)$$

The dynamic error of the ADC (γ_{dyn}) arises due to the delay in information at its output, which leads to the effect similar to the occurrence of the error from the input anti-aliasing filters [1,2].

If we add the dynamic error to error (1) due to the delay, then we can obtain the total error (γ_{tot}) of the ADC in the following form

$$\gamma_{tot} = \frac{1}{2^q} + \gamma_{dyn}. \qquad (2)$$

This error corresponds to the ADC with the effective number of bits.

$$q_{ef} = \left\lfloor \log_2 \left(\frac{1}{\frac{1}{2^q} + \gamma_{dyn}} \right) \right\rfloor. \qquad (3)$$

In (3), the sign $\lfloor\ \rfloor$ means that the nearest smaller integer number is taken.

Effective number of bits q_{ef} is an integrated (qualitative and quantitative) indicator of the information conversion process in the CMS.

II. Research Objective

The main objective and novelty of this article is to obtain the dependence of the effective number of bits on the parameters of the analog-to-digital conversion process for sensors.

At the functional level, the ADC unit has two control sequences (synchronization series) of pulses: select pulses (conversion start pulses) (U_{sel}) and read pulses (output pulses) of the digital code of the converted signal sampling (U_{sam}) [2-16]. The time diagrams of these pulses are presented in Figure 1.

This Figure 1 has the symbols:

t_{pr} – conversion time of the selected sampling into the digital code;

t_{cyc} – period of select pulse feeding.

The research is carried out at the expense of the Grant of the Russian Science Foundation (project № 18-79-10109).

978-1-7281-9899-6 / 20 / $ 31.00 © 2020 IEEE

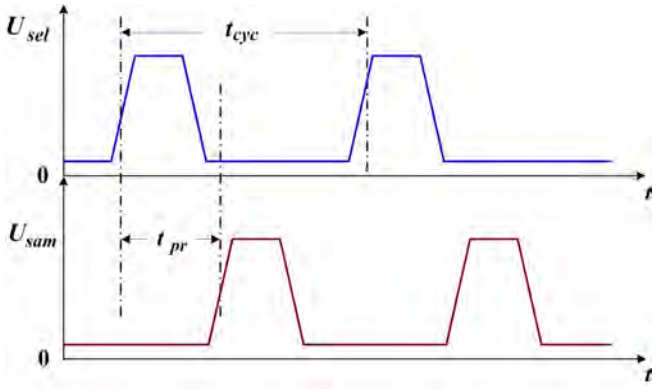


Fig. 1. Timing diagrams of the clock pulses analog-to-digital converter.

For the flash and serial ADCs, the following condition is always met

$$t_{cyc} \geq t_{pr} . \quad (4)$$

For the pipelined ADCs, the situation is possible when

$$t_{cyc} \leq t_{pr} . \quad (5)$$

The value of the maximum relative dynamic error of the ADC for sensors can be represented as [17-23]:

$$\gamma_{dyn} = A_0^{-1} M_{max} \cdot t_{pr}, \quad (6)$$

where: M_{max} – for the transformed waveform, the maximum value of the 1st derivative is;

t_{pr} – information delay period in the analog-to-digital converter;

A_0 – amplitude of the converted signal.

Below, we consider the issues of substantiating the choice of parameters $M_{\max}$, t_{pr} and the sampling frequency ($f_d = \frac{1}{t_{cyc}}$) when finding γ_{dyn} and q_{ef}.

III. Information Delay Time during the Analog-to-Digital Conversion

In general, t_{pr} is the value indicated in the technical documentation for the ADC. But in this issue there may be special situations associated with additional units directly involved in the analog-to-digital conversion process (Fig. 2).

In Fig. 2, the following notation is accepted:
M_{in} – data input module of the CMS;
AS – analog storage;
AM – amplitude multiplexer that selects the desired conversion channel.

At the AM input the AS unit excludes the second-order dynamic error [3] and fixes the input voltage, which is stored in the electric field of its capacitor. At its output the AS maintains a constant voltage at the point of sampling throughout the analog-to-digital conversion.

Most M_{in}s allow installing the ASs at the inputs in the form of additional elements. The AS operation features are related to the protection of the holding capacitor from leakage currents that occur in the aggressive environment, as well as during the unit aging, exposure to humidity and low temperatures.

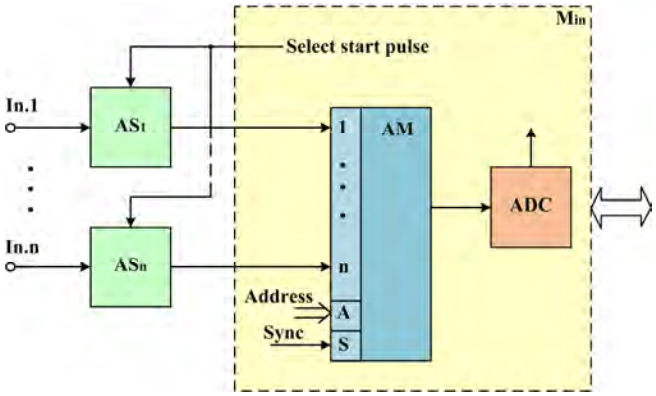


Fig. 2. Functional circuit of the ADC input chain.

If M_{in} is implemented in the technologies of large-scale integrated circuits (LSI) or systems on a chip (SoC), then the ASs are located on the chip. Despite the large areas occupied by the holding capacitors of the AS, this method of eliminating the second-order dynamic error is used in practice of the CMS creating [3,24,25].

In the future, it is assumed that the value of t_{pr} considers the delay for the converted signals in AS, AM.

IV. Determining the Max Amount of the 1st Derivatives from the Converted Signal

In solving this problem, there can be two initial situations determined by the type of SD characteristics of the signal converted in the ADC for sensors:

– the SD characteristics is finite (the signal has a finite spectrum);

– the SD characteristics is infinite (the signal has an extended spectrum).

The signal with the finite spectrum is characterized by the cutoff frequency (ω_s) at the level of $0.707 \cdot A_0$. The maximum 1st derivative of the signal with the finite spectrum (M^{f}_{max}) is determined by the Bernstein inequality [1,3,16]:

$$M^{f}_{max} \leq A_0 \cdot \omega_s . \quad (7)$$

To justify the amount of the maximum 1st derivative of that signal for that extended spectrum, it is suggested to employ the following technique: to describe the signal and obtain the necessary characteristics, draw an analogy of the converted signal with the signal that is received at the output of the Butterworth low-pass filter [16].

An analytical record of the envelope of the SD of the signal with a certain length spectrum, which is obtained at the output of the frequency response of the Butterworth filter, can be represented as:

$$A(\omega) = A_0 \frac{\omega_s^k}{\sqrt{\omega_s^{2k} + \omega^{2k}}}, \quad (8)$$

where k – order of the filter.

The choice of k of the max identity of the spectrum of the original signal with the spectrum of the signal that is obtained at the yield filter is considered in [3].

As shown in [3,16], the maximum 1st-order derivative of the signal with the extended spectrum (M_{max}^{ext}) on the filter yield is defined as:

$$M_{max}^{ext} \le K_s \cdot A_0 \cdot \omega_s, \quad (9)$$

where

$$K_s = \frac{1}{\sqrt{k}} \sqrt[2k]{(k-1)^{(k-1)}}. \quad (10)$$

For a signal at k=1 in compliance with (10):

$$K_s = 1. \quad (11)$$

Further, the article considers the signals with the finite spectrum and with the extended spectrum, the SD of which is identical to the SD of the signal at the output of the first-order Butterworth filter. The maximum derivative for these signals is assigned by the Bernstein inequality (7).

V. Time Sampling Frequency Selection $f_d = t_{cyc}^{-1}$

The processes of time sampling and subsequent reconstruction of signals in the CMS are accompanied by the occurrence of errors. We can define the main errors that determine the choice of f_d:

– spectral foldover error (γ_{fol});

– reconstruction error of method (γ_{rec});

– instrumental errors of sampling and information recovery units.

There is also a well-known sampling theorem [26-31], according to which the sampling frequency of the signal with the finite spectrum is equal to twice cutoff frequency [32-40]:

$$f_d \ge \frac{\omega_s}{\pi}. \quad (12)$$

The selection condition of f_d (12) is boundary and does not take into account the real sampling processes, which are determined by the characteristics of the CMS units. Most often, the choice of the frequency according to (12) is used to obtain margin evaluations of f_d, which are always less (sometimes by several orders of magnitude) than the real sampling frequencies.

In practice, f_d of the input signal of the ADC for sensors is determined by the errors γ_{fol} and γ_{rec}, which are specified when solving the straight error distribution task of the CMS errors. Fig. 3 shows the algorithm proposed in [3] o select the sampling rate of analog signals in terms of the actual work conditions of the question and answer CMS blocks.

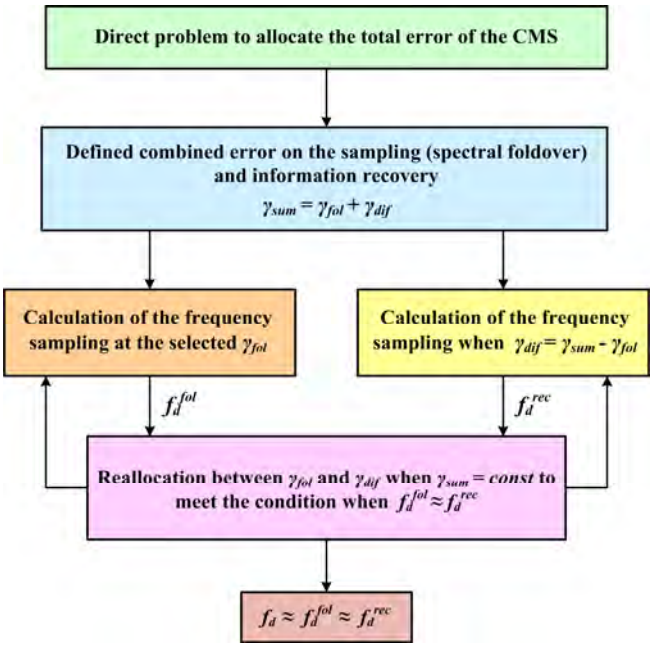


Fig. 3. Algorithm for choosing analog signals in the sample rate in the CMS.

In this Fig. 3:

γ_{sum}– combined error on the sampling and information recovery;

γ_{dif} – the difference between combined error and spectral foldover error;

f_d^{fol} – sampling frequency in terms of γ_{fol};

f_d^{rec} – sampling rate in terms of γ_{rec}.

As can be seen from the algorithm shown in Fig. 3, the sampling frequency is obtained as a result of multiple calculations of f_d^{fol} and f_d^{rec} in the repetitive procedure of the reallocation of errors γ_{fol}, γ_{rec} between each other when their sum is constant.

Then, condition (10) is used to determine f_d, which will give a result more than real during the calculation of q_{ef}. When specifying a specific CMS and calculating the real value of f_d using the iterative algorithm (Fig. 3), it is necessary to correct the value of q_{ef}. It should also be noted, that the proposed estimation algorithm is quite simple and its complexity is low. Due to the lack of other algorithms, complexity comparisons were not carried out.

VI. Calculation of the Effective Number of Bits of the ADC q_{ef}

It was demonstrated above that the dynamic error for the ADC is defined by the formula (6). Then, taking into

account (10), (11), (12), we obtain the max amount for the dynamic error of the ADC in the CMS:

$$\gamma_{dyn} = \pi \frac{t_{pr}}{t_{cyc}}. \quad (13)$$

Substituting the obtained value in (5), we find the total error of the ADC in the following form

$$\gamma_{tot} = \frac{1}{2^q} + \pi \frac{t_{pr}}{t_{cyc}}. \quad (14)$$

This error corresponds to the ADC with the effective number of bits.

$$q_{ef} = \left\lfloor \log_2 \left(\frac{1}{\frac{1}{2^q} + \pi \frac{t_{pr}}{t_{cyc}}} \right) \right\rfloor. \quad (15)$$

Table 1 shows the calculated values of q_{ef} for three ADCs with a number of bits of 14, 10 and 6. It should be noted, that on this topic, the study is fundamental and the only one at the present time, in this regard, a comparison of the results with anything can not be presented.

TABLE I. CALCULATED VALUE OF q_{ef}

$\frac{t_{pr}}{t_{cyc}}$	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-0}	10^{+1}
q_{ef} at $q=14$	13	11	8	4	1	1	1
q_{ef} at $q=10$	9	9	7	4	1	1	1
q_{ef} at $q=6$	5	5	5	4	1	1	1

Fig. 4 shows the graphs plotted in accordance with the data from Table 1.

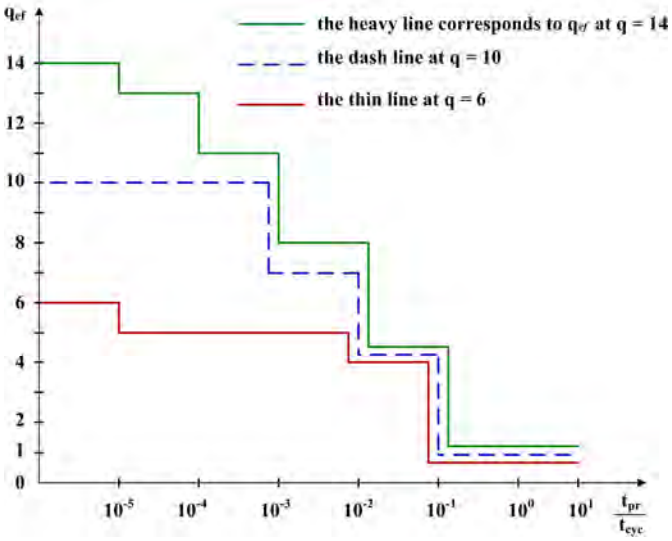


Fig. 4. Plots of changes in the q_{ef} of bits of the ADC depending on the delay time when obtaining the digital readout.

In this Figure, the heavy line corresponds to q_{ef} at q = 14, the dash line at q = 10 and the thin line at q = 6.

VII. CONCLUSION

The effective number of bits of the ADC for sensors depends on the information time delay in it (conversion time), the spectral characteristics of the converted signal and the sampling frequency. The sampling frequency of the signal, calculated with account for the real parameters of the sampling process, significantly exceeds the frequency calculated by the sampling theorem. This requires the correction of the obtained results using the sampling theorem, in the direction of reduction of q_{ef}.

REFERENCES

[1] P.P. Ornatsky, "Fundamentals of informing-measurement technique," 2nd ed., Higher school. Chief publishing house, Kiev, 1983, 455 p. (In Russian)

[2] *L.K. Samoilov,* "Classical Method of The Account of Influence Time Delays of Signals in Devices of Control Systems," Izvestiya SFedU. Engineering sciences, 2016, no. 4, pp. 40-49. (In Russian)

[3] L.K. Samoilov, "Input-output of analog signals in control and monitor systems," Taganrog: Publishing House of South Federal University. 2015. 264 p. (In Russian)

[4] S. Wang, Y. Gao, H. Li, S. Fan, F. Li and L. Geng, "A novel adaptive delay-tracking ADC for DVS power management applications," *2011 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT'2011)*, Beijing, 2011, pp. 65-68, DOI: 10.1109/RFIT.2011.6141791.

[5] S. P. Senapati, "Design of a simple digital ADC using only digital blocks and delay element circuit," *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT'2016)*, Bangalore, 2016, pp. 311-315, DOI: 10.1109/RTEICT.2016.7807833.

[6] S. Sirimasakul and A. Thanachayanont, "A logarithmic level-crossing ADC," 2017 *14th IEEE International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON'2017)*, Phuket, 2017, pp. 576-579, DOI: 10.1109/ECTICon.2017.8096303.

[7] Biao Chen et al., "A Shared-MSB delay-line-based ADC with simultaneous quantization for digital control single-inductor-multiple-output DC-DC converter," *2016 IEEE International Symposium on Integrated Circuits (ISIC'2016),* Singapore, 2016, pp. 1-4, DOI: 10.1109/ISICIR.2016.7829744.

[8] Tsytovich, Leonid I., et al. "Integrating pulse-number ADC with high temporal and temperature stability of characteristics,"in Automatic Control and Computer Sciences 2015, pp. 103-109. DOI: 10.3103/S014641161502008X

[9] T. Wei, W. Liu and L. Yang, "A reference voltage programmable 6-bit differential delay-line ADC for digitally controlled DC-DC switching converters," *2015 IEEE Sixth International Conference on Intelligent Control and Information Processing (ICICIP'2015)*, Wuhan, 2015, pp. 203-208, DOI: 10.1109/ICICIP.2015.7388169.

[10] Azarov O. D. et al., "Static and dynamic characteristics of the self-calibrating multibit ADC analog components," in Optical Fibers and Their Applications 2012, International Society for Optics and Photonics, 2013, pp. 86980.

[11] A. Tritschler, "A Continuous Time Analog-to-Digital Converter With 90µW and 1.8µV/LSB Based on Differential Ring Oscillator Structures," *2007 IEEE International Symposium on Circuits and Systems (ISCAS'2007),* New Orleans, LA, 2007, pp. 1229-1232, DOI: 10.1109/ISCAS.2007.378332.

[12] C. Priyanka and P. Latha, "Design and implementation of time to digital converters," *2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS'2015),* Coimbatore, 2015, pp. 1-4, DOI: 10.1109/ICIIECS.2015.7193116.

[13] C. Belhadj-Yahya, "Analog post amplifier effect on signal sampling and monitoring systems," *2008 15th IEEE International Conference*

on Electronics, Circuits and Systems (ICECS'2008), St. Julien's, 2008, pp. 1253-1256, DOI: 10.1109/ICECS.2008.4675087.

[14] K. Konishi, M. Ishii and H. Kokame, "Stability of extended delayed-feedback control for discrete-time chaotic systems," in *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 46, no. 10, pp. 1285-1288, Oct. 1999, DOI: 10.1109/81.795842.

[15] I. Amri, D. Soudani and M. Benrejeb, "Exponential stability and stabilization of linear systems with time varying delays," *2009 6th IEEE International Multi-Conference on Systems*, Signals and Devices, Djerba, 2009, pp. 1-6, DOI: 10.1109/SSD.2009.4956708.

[16] L. K. Samoylov, "Generalized Bernstein inequality for signals with wide spectrum," Bulletin of Ryazan state Radiotechnical University, 2012, vol. 4, no. 3. (In Russian)

[17] S. Kubo, H. Idei, Y. Tatematsu, T. Saito and M. Iizawa, "Electron Bernstein wave detection by sub-Tera-Hz scattering in the QUEST," 2018 *43rd IEEE International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz'2018),* Nagoya, 2018, pp. 1-2, DOI: 10.1109/IRMMW-THz.2018.8510425.

[18] B. V. Patil, P. S. V. Nataraj and S. Bhartiya, "Application of the Bernstein form for reliable global optimization of mixed-integer problems in process synthesis and design," *2010 2nd IEEE International Conference on Reliability, Safety and Hazard - Risk-Based Technologies and Physics-of-Failure Methods (ICRESH'2010),* Mumbai, 2010, pp. 562-567, DOI: 10.1109/ICRESH.2010.5779611.

[19] H. Igami, *et al.*, "Development of an analysis method on the mode conversion process between electromagnetic and electron Bernstein waves in real experimental configurations," *2009 34th IEEE International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz'2009),* Busan, 2009, pp. 1-2, DOI: 10.1109/ICIMW.2009.5325602.

[20] Y. Gong, "Bernstein filter: A new solver for mean curvature regularized models," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2016*), Shanghai, 2016, pp. 1701-1705, DOI: 10.1109/ICASSP.2016.7471967.

[21] W. Huazhang, D. Jiajia and Z. Tingting, "Bernstein Bezout matrix with applications to polynomial stability," *2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC'2017),* Hefei, 2017, pp. 639-644, DOI: 10.1109/YAC.2017.7967488.

[22] G. M. Phillips, "A survey of results on the q-Bernstein polynomials," in IMA Journal of Numerical Analysis, vol. 30, no. 1, pp. 277-288, Jan. 2010, DOI: 10.1093/imanum/drn088.

[23] J. Huang and S. P. Kuo, "Parametric excitation of electron Bernstein waves through a thermal oscillating two stream instability," *IEEE International Conference on Plasma Sciences (ICOPS'1993),* Vancouver, BC, Canada, 1993, pp. 92, DOI: 10.1109/PLASMA.1993.593064.

[24] A. Polishchuk, "Programmable analog ICs Anadigm: structures and principles of construction," Journal Modern Electronics, no. 1, 2005 pp. 24-27. (in Russian)

[25] P.P. Redkin, "Precision data collection systems of the MSC12xx family of Texas Instruments: architecture, programming, application development," Dodeka-XXI Publishing House, Moscow, 2006, 608 p. (in Russian)

[26] Jerri, Abdul, "The Shannon Sampling Theorem—Its Various Extensions and Applications: A Tutorial Review," in Proceedings of the IEEE. Nov. 1977, vol. 65 (11), pp. 1565–1596, DOI:10.1109/proc.1977.10771

[27] M. Unser and J. Zerubia, "A generalized sampling theory without band-limiting constraints," in *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 8, pp. 959-969, Aug. 1998, DOI: 10.1109/82.718806

[28] Qiao Wang and Lenan Wu, "A sampling theorem associated with quasi-Fourier transform," in *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 895-, March 2000, DOI: 10.1109/78.824688.

[29] R. S. Stankovic and J. Astola, "Reading the Sampling Theorem in Multiple-Valued Logic: A Journey from the (Shannong) Sampling Theorem to the Shannon Decomposition Rule," *37th IEEE International Symposium on Multiple-Valued Logic (ISMVL'2007),* Oslo, 2007, pp. 2-2, DOI: 10.1109/ISMVL.2007.48.

[30] H. Ueda and T. Tsuboi, "A sampling theorem for periodic functions with no minus frequency component and its application," *2013 19th Asia-Pacific Conference on Communications (APCC'2013),* Denpasar, 2013, pp. 225-230, DOI: 10.1109/APCC.2013.6765946.

[31] J. Long, P. Ye and X. Yuan, "Truncation error and aliasing error for Whittaker-Shannon sampling expansion," Proceedings of the 30th Chinese Control Conference, Yantai, 2011, pp. 2983-2985.

[32] J. L. Brown, "Summation of certain series using the Shannon sampling theorem," in *IEEE Transactions on Education*, vol. 33, no. 4, p. 337-340, Nov. 1990, DOI: 10.1109/13.61086.

[33] V. A. Kotel'nikov, "On the transmission capacity of 'ether' and wire in electric communication," J. Uspekhi Fizicheskikh Nauk and Russian DOI: 10.1070/PU2006v049n07ABEH006160 (In Russian)

[34] K. Shannon, "Works on Information Theory and Cybernetics," Publishing house of foreign literature, Moscow, 1963, 824 p. (In Russian)

[35] Academy of Sciences, 2006, vol. 49 no. 7. pp. 762-770. V. V. Zamaruiev, "The use of Kotelnikov-Nyquist-Shannon sampling theorem for designing of digital control system for a power converter," *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON'2017),* Kiev, 2017, pp. 522-527, doi: 10.1109/UKRCON.2017.8100305.

[36] Guo Tiande and Gao Ziyou, "A sampling theorem without band-limiting constraints," *2000 5th IEEE International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000 (WCC'2000 – ICSP'2000.)*, Beijing, China, 2000, vol. 1, pp. 89-94, DOI: 10.1109/ICOSP.2000.894451.

[37] C. Fei-na and L. Qin-xian, "Fourier Transform and Reconstruction of Periodic Signal Based on Non-Uniformity Sampling," 2007 *IEEE International Conference on Control and Automation (ICCA'2007),* Guangzhou, 2007, pp. 2581-2583, DOI: 10.1109/ICCA.2007.4376828.

[38] X. Liu, W. Li, J. Wei and L. Cheng, "Adaptable Hybrid Filter Bank Analog-to-Digital Converters for Simplifying Wideband Receivers," in *IEEE Communications Letters*, vol. 21, no. 7, pp. 1525-1528, July 2017. DOI: 10.1109/LCOMM.2017.2690281

[39] A. Anis, A. Gadde and A. Ortega, "Towards a sampling theorem for signals on arbitrary graphs," 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2014),* Florence, 2014, pp. 3864-3868, DOI: 10.1109/ICASSP.2014.6854325.

[40] H. Fraz, N. Bjorsell, J. S. Kenney and R. Sperlich, "Prediction of Harmonic Distortion in ADCs using dynamic Integral Non-Linearity model," 2009 *IEEE Behavioral Modeling and Simulation Workshop (BMAS'2009),* San Jose, CA, 2009, pp. 102-107, DOI: 10.1109/BMAS.2009.5338881.

Coupled Piecewise Constant Memristor based Reactance-less Oscillators

Vladimir V. Rakitin,
IPPM, Russian Academy of Sciences
Moscow, Russia
vlarak@rambler.ru

Sergey G. Rusakov, *Member IEEE*
IPPM, Russian Academy of Sciences
Moscow, Russia
rusakov@ippm.ru

***Abstract*— New type of memristor-based reactance-less oscillators circuit is considered – memristor piecewise constant oscillators. The phase plane methodology was applied for analysis of coupled memristor-based reactance-less piecewise constant oscillators (PWC MBO) under external excitation. The capabilities of PWC MBO are discussed. The results of oscillator simulation confirm these capabilities.**

Keywords—memristor devices, piecewise-constant oscillators, reactance-less memristor based oscillators, two-threshold comparator, phase plane.

I. INTRODUCTION

Artificial neurons (AN) are widely used as the main elements in promising neuromorphic systems [1]. Several generations of AN models have changed for period of their development. In particular, the modern oscillatory artificial neurons provide quite complex responses to external signals. In fact, they are generators of binary pulse sequences with complex modulation system that depends on both the intensity and polarity of the incoming signals, as well as on the previous time events. Various electronic devices can be used for their implementation [2].

Recently AN models based on piecewise constant (PWC) oscillators have appeared [3-5]. They are developed on the base of standard electronic components including amplifiers, logic gates, resistors, capacitors. The transient processes occur in these circuits under constant excitation, for example the charge or discharge of the capacitor at constant current.

Thus piecewise constant (PWC) oscillators are the oscillators with mathematical models which are systems of ordinary differential equations (ODE) with piecewise constant coefficients. The piecewise linear functions of time are the solutions of such ODE systems. Correspondingly the signals generated by AN in this case are piecewise linear functions of time.

The features of such circuits simplify the analysis. At the same time, such circuits preserve the diversity of their behavior. They also provide the properties of more complex models of AN including excitation, inhibition, pulse generation and pulse trains generation. The analysis of the piecewise constant PWC oscillators and AN based on these oscillators is given in a number of works [6-8].

As a rule, PWC oscillators are based on linear integrators controlled by nonlinear elements with binary outputs, such as comparators and hysteresis comparators. The simplest PWC oscillator can contain only one integrator. But its capabilities to modulate pulse sequences are limited in this case. Therefore, at least two integrators in PWC oscillators are used to construct AN circuit. These two integrators specify the variables in PWC oscillator model. The linear superposition of these variables is input signal for nonlinear control elements. Such connected PWC oscillators can be considered as coupled dynamical system. The complex behavior of this system requires the complicate analysis and synthesis but provides a variety of PWC oscillator functionalities.

This paper is devoted to application of memristor devices as integrating elements of the artificial neural circuits. The properties of memristors [9] open up new possibilities of constructing the memristor generators and artificial neurons on their base [10].

The inertial property of memristors provides the elimination of reactive elements (inductors and capacitors) in generator circuits. As a result, reactance less memristor based oscillators can be constructed [11-13] with the characteristics desired for AN generation. This type of oscillators can operate in the frequency range from kHz to MHz with power consumption varying in the range from nanowatt to microwatt.

The nonlinearity of the memristor characteristics due to the change in its resistance when current flows through device limits the development of PWC memristor based oscillators (PWC MBO). However, if only the current sign via the memristor is changed, then this restriction is removed. To control the generation process it is enough to send the input signal not to the memristor, but to the nonlinear element. This circuit property can be implemented in memristor piecewise constant generators.

The purpose of this paper is to consider the behavior of such memristor piecewise constant generators and their opportunities for application in AN.

The simplest PWC MBO contains a single memristor. The connection of two such generators provides new capabilities of pulse train modulation that is required for AN oscillator circuit development.

The rest of the paper is organized as follows. Section 2 presents the basic principles of PWC MBO constructing. PWC MBO circuit and its functionalities are discussed in section 3. In section 4 the analysis of two coupled PWC MBO is performed. Some simulation results of coupled PWC MBO are presented in section 5.

The reported study was funded by RFBR, project number 19-29-03012.

II. The Principles of Construction of Piecewise Constant Memristor based Oscillators

To provide the occurrence of relaxation oscillations in considered circuit type the connection of a memristor with a nonlinear active element must meet to certain requirements [13]. In particular, the two-threshold comparator (TTC) [12] meets these requirements. In memristor based generator (MBO) the current flow changes the resistance value of the memristor. The corresponding change in the voltage on the memristor causes TTC switching that provides a periodic change in the direction of the current.

The TTC output signal can be used to transmit binary signals ("0" and "1") within the network of connected MBO. Usually the input signal in MBO changes the current through the memristor and controls the rate of the memristor state change. This process can be considered as integration of input signal. When control signals are received the time change in the voltage on the memristor has nonlinear character. Thus, MBO with traditional control is not be piecewise constant generator.

The main idea to eliminate the time nonlinearity is to control MBO not by the input current amplitude but by its transmission time. To achieve this goal, it is suggested to send the input signal not directly to the memristor device but to the TTC circuit. In this case the control by TTC is reduced to change in its threshold voltages.

Increasing the maximal threshold voltage and decreasing the minimal threshold voltage leads to increasing the possible range of memristor voltage variation. Respectively decreasing the maximal threshold voltage and increasing the minimal threshold voltage leads to reducing the possible range of memristor voltage variation. Accordingly, the possible range of changes in the resistance of the memristor will also change. After completion input signal activity the state of MBO may also change. This depends on time relations between oscillations in MBO and time duration of signal at MBO input.

It is important that all these changes occur under constant value of the current through the memristor. By such a way MBO becomes the piecewise constant generator (PWC MBO).

Interconnection of the piecewise constant generators (PWC MBO).

The system of connected PWC MBO can be considered. In particular, the circuit involving transmitting PWC MBO and receiving PWC MBO can be discussed.

The binary output of PWC MBO can control the thresholds of other such generators in various ways. For example, high output level ("'1") of the transmitting PWC MBO increases one or both TTC thresholds of the receiving PWC MBO while a low level ("0") causes the opposite effect.

Below we will limit ourselves by consideration of the following algorithm of PWC MBO interconnect behavior:

- the direct output high level ("1") at output of transmitting PWC MBO causes reducing the difference between the TTC thresholds values and it causes an increase in the difference between the threshold values in inverted version;

- the output low level ("0") of the transmitting PWC MBO does not impact on the TDC threshold of the receiving PWC MBO;

- change in thresholds is small enough so that the generation of the receiving PWC MBO is preserved.

III. Memristor Piecewise Constant Generator

The following PWC MBO circuit is proposed for further consideration (Fig.1).

The sheme of piecewise constant memristor based oscillator with controlled threshold parameters (PWC MBO) consists of memristor M and a two-threshold comparator (TTC). The comparator output is connected to the memristor M using the current source I_M. The TTC is formed by two comparators whose outputs are combined by a logical NAND circuit. The output binary signal V_{OUT} ("0" and "1") is converted into currents $-I$ or $+I$ using current generator I_M. These currents provide the reversible change in the resistance of the memristor. The voltage from the memristor v and the input signal V_{IN} are received at TTC inputs. Moreover the voltage V_m and V_M set the initial minimal and maximal threshold voltages, respectively. Note that for $V_{IN} > V_m$ PWC MBO output state is set to logical "0".

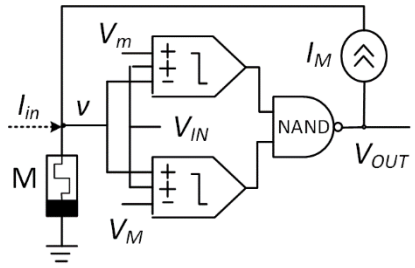


Fig. 1. Schematics of memristor based piecewise constant oscillator (PWC MBO). The current input I_{in} (dotted line) demonstrates the traditional method of MBO control.

Two variables specify the state of PWC MBO circuit: the resistance value of the memristor $R(t)$ and TTC output voltage V_{OUT} . The differential equation $dR/dt = -\gamma \cdot I_M$ takes place when connecting memristor anode to the current generator I_M (Fig. 1). Here coefficient γ determines the switching speed of the memristor device. The value of γ is constant in the framework of the drift-diffusion memristor model [9].

Below we will go to dimensionless variables. In this case we get I=1, γ=1, and as a result the simple form of differential equation: $dR/dt = \pm 1$.

The variation of memristor resistance R is limited by minimal R_{ON} and maximal R_{OFF} values. The range of variation R is additionally narrowed due to PWC MBO circuitry with the maximal V_M and minimal V_m values of TTC threshold voltages and also the input signal $V_{IN}(t)$:

$$R_{ON} < (V_m + V_{IN}(t))/I = R_m + r_{IN}(t) \le R(t), \quad (1a)$$

$$R(t) \le R_M - r_{IN}(t) = (V_M - V_{IN}(t))/I < R_{OFF}. \quad (1b)$$

Here $R_m = V_m/I$ is the minimal threshold resistance, $R_M = V_M/I$ - the maximal threshold resistance, $r_{IN}(t) = V_{IN}(t)/I$ – the change in the threshold resistance due to the input signal.

Thus the process of generation in PWC MBO reduced to change $R(t)$ in the range of threshold resistances. The additional requirement is introduced to maintain nonzero range of memristor resistance changes and prevent going beyond the linear range.

$|r_{IN}(t)|min\{R_m - R_{ON}, R_{OFF} - R_M, (R_M - R_m)/2\}$ (2).

Input PWC MBO does not directly affect the resistance change. It impacts on the switching threshold, i.e. it affects the switching time points with change the direction of resistance variation. Thus, the ability to change the state and output signal of PWC MBO depends on the time of arrival of the control signal. Note that when the control signal is applied to the memristor, the change in the state of the memristor in MBO does not depend on the time of arrival of the control signal.

The action of the input signal *IN* on both types of generator (V_{IN} on PWC MBO or I_{in} on MBO) is shown in Fig.2. As we can see from the figure the MBO response depends on the duration of the input signal, and the PWC MBO response depends on the time point of the input signal arrival. The first two broad pulses speed up and slow down transient processes in MBO but do not impact on PWC MBO, since they operate outside of TTC switching intervals. Two subsequent short pulses have a weak effect on the MBO, but lead to a change in the behavior of the PWC MBO, as they change its switching thresholds.

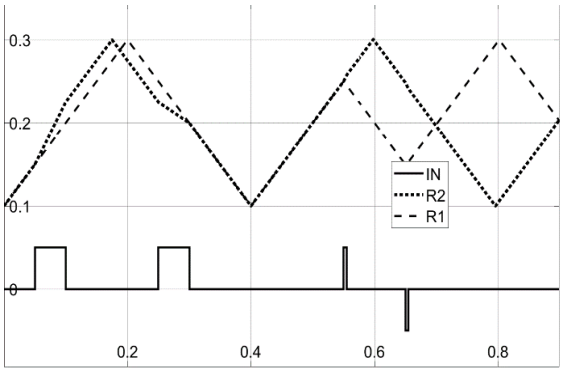


Fig. 2. Comparison of impact of input signal (solid line) on PWC MBO (dashed) and MBO (dotted line).

PWC MBO can be used as a signal source for another-receiving PWC MBO.

Under the positive pulse at input of receiving PWC MBO the transmitting PWC MBO locks the phase of the receiving one. The difference between the thresholds of the receiving PWC MBO is reduced under logical "1" at output of the transmitting PWC MBO. If the receiving PWC MBO is lagging in phase (its positive output signal is lagging), the maximum threshold is lowered, the lag is reduced and the phase lag is reduced. If the receiving PWC MBO is ahead of the phase (its positive output signal appears earlier) If the receiving PWC MBO is ahead of the phase (its positive output signal appears earlier), then the minimal threshold will be decreased. In this case the transition time to the lower zero level will be delayed and the advance will be reduced.

The transmitting PWC MBO imposes an antiphase state on the receiving one under negative input pulse.

IV. Analysis of Two Coupled Memristor Piecewise Constant Generators (PWC MBO)

Two coupled PWC MBO behave in more complex way. Their joint behavior depends significantly on the coupling strength factor and the connection direction.

The system of opposite connected PWC MBO1 and PWC MBO2 with binary output signals is shown in Fig.3. The input of PWC MBO1 receives an inverted signal $(-k_1V_2)$ and analog drive signal V_C. The input of PWC MBO2 receives direct signal k_2V_1. The system contains a phase detector $F(V_1, V_2)$ that performs logical function over the binary outputs PWC MBO1 and PWC MBO2. For example, if the function is "exclusive OR" then zero output signal $V_S = 0$ corresponds to completely in-phase input signal PWC MBO1 and PWC MBO2.

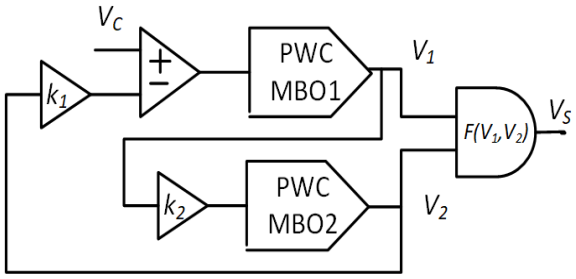


Fig. 3. The system of coupled PWC MBO with the input adder and the output phase detector.

If the PWC MBO generators are the same then the modules of the rates of change of their resistances R_1 and R_2 are also the same. But direction signs of the rates may differ. Thus the system state is described by the variables R_1 and R_2 and direction signs of dR_1/dt and dR_2/dt. The behavior of such a system can be conveniently considered at the phase plane of variables $R_1 \div R_2$. Since the derivatives of variables take the values ± 1, the trajectory of the image point is inclined straight lines (then inclined) with angle of $\pm \pi/4$ relatively the coordinate axes (see Fig.4 and Fig. 5.).

Thus one from the four trajectory defined by the signs of dR/dt can pass through each point of the phase plane. The own boundary corresponds to each of them. The trajectory of the image point is mirrored from the boundary after reaching boundary. The image point moves along the trajectory in the opposite direction when it hits the corner of the rectangle formed by the borders.

Application of phase plane for analysis.

The phase portrait of the system depends qualitatively on the ratio of coupling coefficients k_1 and k_2:

$$r_1 = -k_1V_2(t)/I, \quad (3)$$

$$r_2 = k_2V_1(t)/I.$$

a) The case of equal values ($|r_1| = |r_2| = r$) is a degenerate case.

Two squares can be distinguished on the phase plane: the large one with the side $R_M - R_m + r$ and the small one with the side $R_M - R_m - r$ (Fig. 4).

The performed analysis showed:

- any rectangle formed by inclined lines, inscribed in a small square, is a stable trajectory of movement counterclockwise. The diagonals of a small square are also stable trajectories;

- any rectangle with sides larger than $2r\sqrt{2}$, formed by inclined lines, inscribed in a large square, is a stable clockwise movement trajectory;

- all other inclined lines correspond to unstable trajectories that turn into stable ones;

The phase portrait in Fig.4 corresponds to the bifurcation of the system. The character of its behavior changes qualitatively when the values r_1 and r_2 are differ.

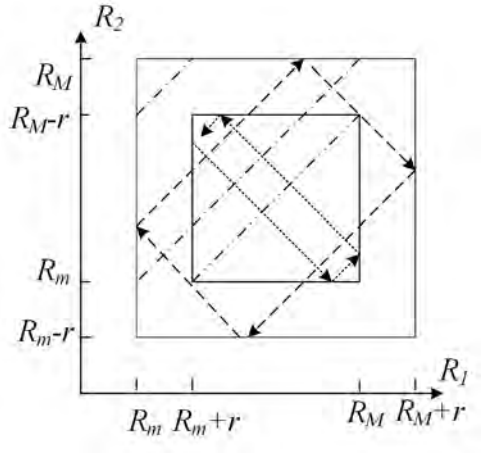


Fig. 4. Phase plane of opposite connected PWC MBO for degenerate case.

b) The case of different values ($r_1 \neq r_2 = r$)

There are two possible versions:

$|r_1| = r_m < r = r_2$, $|r_1| = r_M > r = r_2$.

In the first case we can select the parallelogram with opposite vertices $(R_m + r,\ R_M - r_m)$ and (R_M, R_m) at the phase plane (Fig.5). This parallelogram is formed by inclined lines with angle (-π/4) (dashed lines in Fig.6) and lines that are parallel to the abscissa axis. Inclined lines with angle (-π/4) lying inside this parallelogram are stable trajectories. They correspond to antiphase oscillations with the period

$T = 2\,(R_M - R_m - r_m)/\gamma I$.

In the second case the parallelogram with opposite vertices $(R_M,\ R_M - r_M)$ and $(R_m + r,\ R_m)$ can be selected that is formed by inclined lines with angle (π/4) and lines paralleling to the ordinate axis. The inclined lines with angle (π/4) lying inside this parallelogram are stable trajectories. They correspond to in-phase oscillations with the period

$T = 2\,(R_M - R_m - r)/\gamma I$.

Any other trajectory becomes stable during the certain number of cycles. So for the first case the trajectory starting at point *A* after several switches transfers to the stable trajectory with a negative slope (Fig.5).

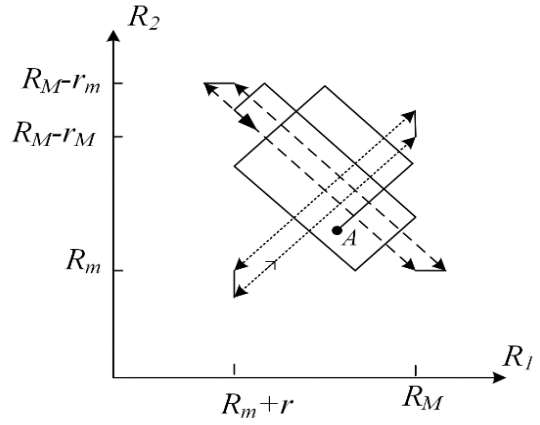


Fig. 5. Phase plane of opposite connected PWC MBO with strong positive coupling (dotted line) and a strong negative coupling (dashed line).

Techniques to drive the system

The drive signal can change the phase trajectory of the system. Let the case be $r_1 = r_m$.

An additional external drive signal V_C leads to shift in the range of varying PWC MBO1 threshold voltage. Therefore the additional shift in the threshold resistances $r_C = V_C/I$ is generated. As a result, in accordance with (1) the following restrictions are active for PWC MBO1 and PWC MBO2 during the action of the V_C signal:

$$R_m + r - r_C \le R_1(t) \le R_M + r - r_C \quad at\ \ dR_2/dt < 0, \tag{4}$$

$$R_m - r_C \le R_1(t) \le R_M - r_C \ \ at\ \ dR_2/dt > 0\,, \tag{5}$$

$$R_m - r \le R_2(t) \le R_M - \ r \ \ at\ \ dR_1/dt < 0, \tag{6}$$

$$R_m \le R_2(t) \le R_M \quad\ \ at\ \ dR_1/dt > 0\,. \tag{7}$$

The phase portrait of the system with drive signal is presented in Fig.6.

Let the system initially be in a stable state corresponding to trajectory *ab*. The drive signal V_C corresponds to trajectory *(gh*). The transfer to trajectory *(gh*) is unavoidable under constant drive signal of quite long activity. This is result of change of dR_2/dt sign and reducing the threshold resistance (4) to $R_m + r - r_C$ when the point *b* reaches the border at $R_1(t) = R_m + r_m$. As consequence the image point moves from *b* to *c*, then to *d*, until it falls on the trajectory (*gh)*.

The speed of the transition process to steady state depends on the difference $\Delta = r_2 - r_1$. The width of the stability area also depends on it. The return to stationary trajectory after external excitation may be long at low values Δ. The proportional dependence of the return time on the amplitude of the external signal can be expected due to the piecewise linear character of transient waveforms in certain range of its amplitude variation.

The behavior of self-oscillating coupled PWC MBO is described by piecewise constant differential equations. In this case the complete analytical solution can be obtained. For this goal problem of elastic reflection of a point from the sides of a rectangle is solved. The problem is solved in version when the position of the rectangle sides depends on the sign of the speed of point movement.

In general case it is expedient to solve the considered problems using simulators in particular for simulation of long time periods and simulation of coupled PWC MBO with external signals.

V. Simulation of Behaviour of Coupled Memristor Piecewise Constant Generators (PWC MBO)

Below some simulation results of the PWC MBO circuits versions are presented.

The dimensionless parameters and variables are used below: $R_{OFF} = 1$, $R_M = 0.6$, $R_m = 0.2$, $R_{ON} = 0.1$, V=1, I=1, $\gamma = 1$. The dimensionless time is also applied.

The example of the behavior of the system with phase detector NOR and various values of coupling factors ($r_2 = 0.1,\ r_m = 0.09$) is shown in Fig. 7.

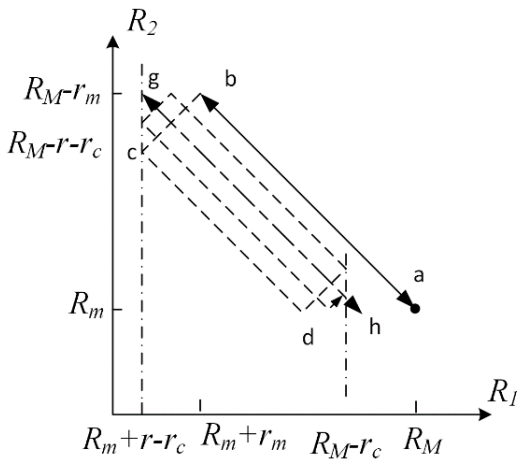


Fig. 6. Example of a phase plane of system of coupled PWC MBO under control.

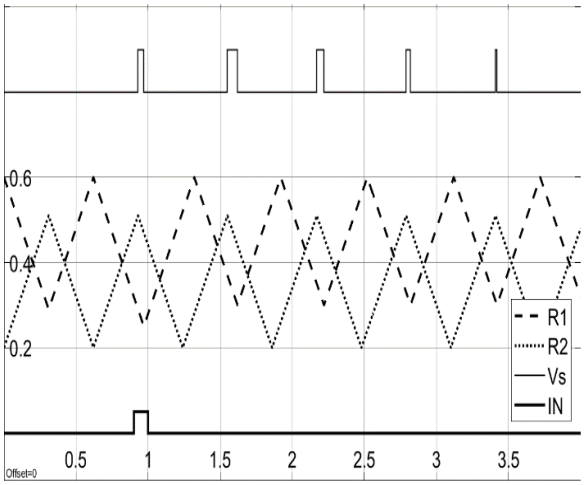


Fig. 7. The computed waveforms in the system of coupled PWC MBO with short input signal.

Initially the behavior of the system corresponds to stable trajectory with antiphase oscillations. At time *t=0.9* the positive pulse arrives which has the amplitude of *0.05* and duration of *0.1*. It delays the switching of PWC MBO1. The system goes into an excited state and the antiphase is violated during four periods of oscillation. This leads to series of five output pulses. The similar behavior occurs if a negative input pulse is applied instead of a positive one, for example, at time *t=0.6*.

The behavior of the system becomes more complex if the pulse duration increases. If the pulse duration exceeds the transition time to a stable trajectory, the number of output pulses is doubled

The speed of transition to a stable trajectory depends significantly on the difference Δ in the coupling factors. If this value is small then the process can be delayed. The time of transition to a stable trajectory at fixed values of connections depends on the amplitude and time of arrival of the input pulse. This can be used to perform the conversion of the external excitation value to the duration of the transition process.

The carried out analysis and the simulation results show that the system of two identical opposite connected PWC MBO with a phase detector has the following properties:

1) the stable oscillations are antiphase or in-phase in self-oscillating mode if the difference in the coupling coefficients is positive or negative, respectively;

2) when the drive signal changes, the transition to a new stable state occurs only if the pulse amplitude is sufficient to change the threshold value during the pulse activity. It is accompanied by violation of oscillation synchronization, which leads to the appearance of pulses at the detector output. The duration of the transition process and the number of pulses at the detector output are inversely proportional to the modulus of the difference in the coupling coefficients and proportional to the amplitude of the drive signal.

3) when the duration of the input pulse exceeds the transition process the number of output pulses is doubled.

VI. Conclusion

The principles of operation of new class of memristor based piecewise constant oscillators were presented. It was shown that the system of two opposite connected memristor generators PWC MBO has wide functionality for converting analog and analog-digital signals into binary pulse sequences and is suitable for creating simple models of generating AN.

References

[1] C. Schuman, T Potok., R. Patton, J. Birdwell, M. Dean, G.Rose, J. Plank, "A Survey of Neuromorphic Computing and Neural Networks in Hardware." arXiv:1705.06963, V1, 19 May 2017.

[2] R.Islam et al, "Device and Materials Requirements for Neuromorphic Computing", Journal of Physics D: Applied Physics, V.52, N.11, 2019.

[3] T. Tsubone, T. Saito, "Manifold piecewise constant systems and chaos," IEICE Trans. Fundamentals, E82-A, N 8, pp.1619-1626, 1999.

[4] C. Matsuda, H. Torikai, "A Novel Generalized PWC Neuron Model: Theoretical Analyses and Efficient Design of Bifurcation Mechanisms of Bursting," IEEE Transactions On Circuits and Systems II: Express Briefs V. 11, N. 4, 2012.

[5] Y. Yamashita., H. Torikai, "Theoretical Analysis for Efficient Design of a Piecewise Constant Spiking Neuron Model," IEEE Transactions on Circuits And Systems II: Express Briefs, V. 61, N.1, pp 54-58, 2014.

[6] K. Mitsubori, T. Saito, "Dependent Switched Capacitor Chaos Generator And Its Synchronization," IEEE Transactions On Circuits And Systems. I, V. 44, N. 12, pp. 1122–1128, 1997.

[7] Y. Matsuoka, "Master–Slave Coupled Piecewise Constant Spiking Oscillators," IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 94-A(9): 2011. pp.1860-1863.

[8] T. Tsubone, T. Saito, N. Inaba, "Design of an analog chaos-generating circuit using piecewise-constant dynamics", Prog. Theor. Exp. Phys., 053A01, 2016.

[9] D. Strukov, G. Snider., D. Stewart, "The missing memristor found," Nature (London, U.K.), V. 453, pp. 80–83, 2008.

[10] Handbook of Memristor Networks (Editors Chua L., Sirakoulis G., Adamatzky A.), Springer, 2019.

[11] M. Zidan, et al. H. Omran, C. Smith, A.G. Radwan, K.N. Salama, "A Family of Memristor Based Reactance-Less Oscillators," Int. J. Circuit Theory and Applications. V. 42. № 11. pp. 1103–1122, 2013.

[12] A.G. Radwan., M.E. Fouda, "On the Mathematical Modeling of Memristor, Memcapacitor, and Meminductor", Cham: Springer International Publishing Switzerland, 2015.

[13] V.V. Rakitin., S.G Rusakov, "Operating principles of reactance-less memristor-based oscillators," J. Commun. Technol. Electron., 2017, vol. 62, no. 6, pp. 621–625.

Unidirectional Emission of Active Eccentric Microring Cavities

Anna I. Repina
Department of System Analysis and Information Technologies
Kazan Federal University
Kazan, Russia
airepinas@gmail.com

Alina O. Oktyabrskaya
Institute of Computational Mathematics and Information Technologies
Kazan Federal University
Kazan, Russia
alina.oktyabrskaya.21@gmail.com

Evgenii M. Karchevskii
Department of Applied Mathematics
Kazan Federal University
Kazan, Russia
ekarchev70@gmail.com

***Abstract*— Calculations of spectra, thresholds, and emission directivities of laser modes for active eccentric microring cavities are based on the lasing eigenvalue problem, which was reduced to the system of Muller boundary integral equations, and was discretized by the Galerkin method with trigonometric basis. In computational experiments, we demonstrate unidirectional emission of lasing, which is achievable with some assumptions on the position of the air hole in the cavity and its radius. In this work we obtain high directivity for unidirectional emission with the one pronounced ray in the directional pattern.**

Keywords—microring laser, active microcavity, lasing eigenvalue problem

I. INTRODUCTION

For more than few decades, optical microdiscs and microrings [1,2] are growing their popularity in scientists society. Due to their shape (thin and flat) they have interesting properties for investigation [3,4]. For this case, the special formulation was developed for lasing modes analysis [5]. Lasing Eigenvalue Problem (LEP) for the two-dimensional formulation is the well-known physical model [6,7,8]. Thresholds of lasing and emission frequencies can be considered as solutions of LEP for such lasers as in [5,9]. This statement was used for calculation of characteristics of lasers with active zones of various configurations [10-15]. In [16-19], it was shown how to use the LEP formulation for nanolasers with surface-plasmon modes.

In [20], the unidirectional lasing emission for eccentric microring cavities was demonstrated experimentally. In this paper, we investigate LEP for the purpose of modelling the unidirectional emission of lasing numerically. With the help of LEP we can detect the directivities of lasing and the threshold values of the gain and study spectrum properties [9-15, 21-24, 26-28], also the disclosure of unidirectional emission is becoming achievable. In this work we use the method proposed in [13] and further developed in [14] to study unidirectional emission of lasing for higher azimuthal order modes. We investigate the unidirectional emission of lasing for several modes of an active eccentric microring cavity, which is controlled by position and size of the hole. It is important to notice that the unidirectional emission can be achieved only for even solutions of LEP. This structure is important in the case of the development of active microcavities.

II. LASING EIGENVALUE PROBLEM

Let us formulate LEP following [10] for an active eccentric microring cavity, which geometry is shown at Fig. 1 (see also [25] for analogous statements for waveguide problems). The considered domains Ω_1, Ω_2, and Ω_3 are separated by the boundaries Γ_1, Γ_2 and have the refractive indices denoted by $\nu_m, m = 1,2,3$. In the annular region Ω_2, the refractive index is a complex value $\nu_2 = \alpha_2 - i\gamma$, where γ is the positive parameter of the active domain. In the exterior unbounded domain Ω_3, the refractive index equals to the same value of the interior region, $\nu_1 = \nu_3 = \alpha_1$. Here, $\alpha_1,\ \alpha_2 > 0$ are known real parts of ν_m, while the values of the gain index $\gamma > 0$ are expected to be found.

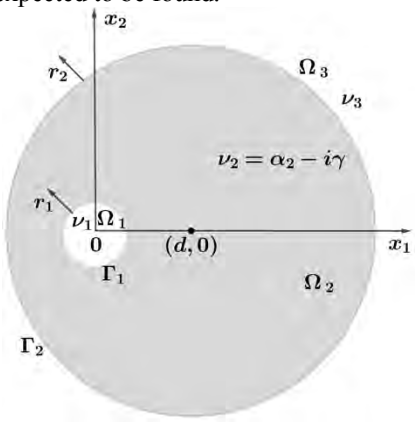


Fig. 1: Geometry of an eccentric microring cavity

Let us find solutions of LEP as pairs of the positive eigenvalues (k, γ). We denote by u the eigenfunction corresponding to the wavenumber k and the threshold gain γ. Then the following relations have to be satisfied [9]:

$$\Delta u + k_m^2 u = 0,\ x \in \Omega_m,\ m = 1,2,3, \qquad (1)$$

$$u^- = u^+,\ \eta_m \frac{\partial u^-}{\partial r_m} = \eta_{m+1} \frac{\partial u^+}{\partial r_m},\ x \in \Gamma_m,\ m = 1,2, \qquad (2)$$

$$\left(ik_3 - \frac{\partial}{\partial r}\right) u = o\left(\frac{1}{\sqrt{r}}\right),\ r = |x| \to \infty. \qquad (3)$$

Here, $u = H_3\ (E_3)$, $\eta_m = \nu_m^{(-2)}(1)$, $m = 1,2,3$, in the case of H-polarized (E-polarized) mode, as usual, $k_m = k\nu_m$.

III. NUMERICAL RESULTS

In our computations, we are looking for the H-polarized modes of the eccentric microring cavity with the refractive index in the hole and in the environment equal to $\nu_1 = \nu_3 = 1$. The real part of ν_2 in the bounded area Ω_2 is $\alpha_2 = 2.63$. We are looking for unidirectional emission of lasing in the assumption that the relative radius of the hole equals 0.15. We place the air hole in the center, after that we start to shift it to the boundary Γ_2.

At Fig. 2, we see the dependencies of the gain threshold on the emission frequency for the fixed relative distance d between the centers of the circles. The initial values for the centered hole (for $d = 0$) are marked by triangles and circles for even and odd solutions, respectively. The solutions for even and odd modes for fixed distance $d = 0.590$ are marked

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

by stars, which related to the maximum directivity D of the 13th mode. We calculate the directivity D following [9]. We denote the normalized frequency of lasing by $\kappa = ka_2$, where a_2 is the external radius of Ω_2.

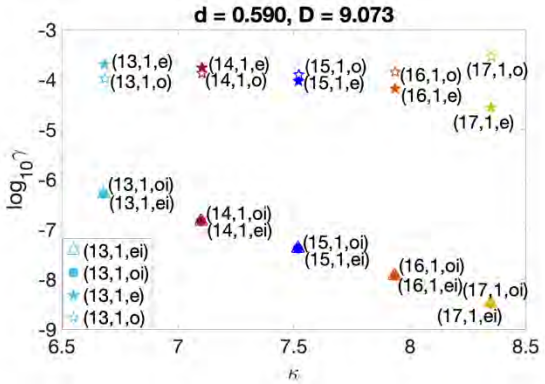


Fig. 2: Values of κ and $\log_{10}\gamma$ for $d = 0.590$, the maximum directivity of the $H_{13,1,e}$ mode equals $D = 9.073$

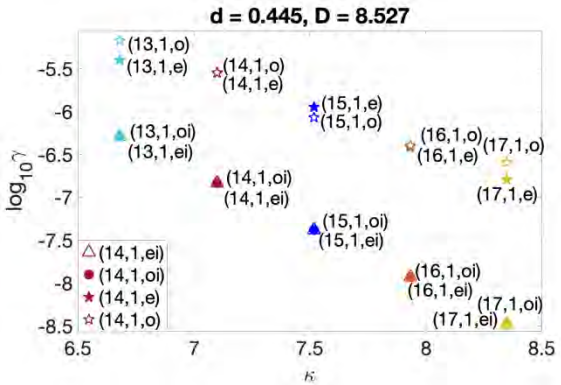


Fig. 3: Values of κ and $\log_{10}\gamma$ for $d = 0.445$, the maximum directivity of the $H_{14,1,e}$ mode equals $D = 8.527$

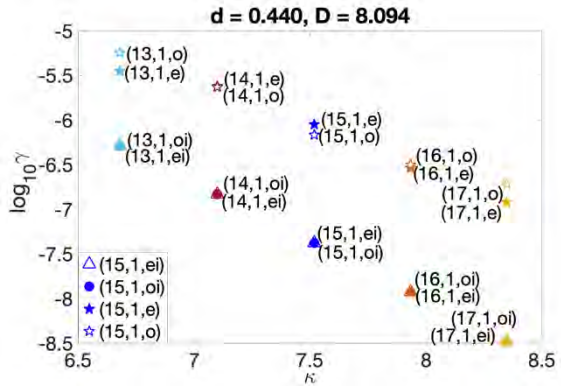


Fig. 4: Values of κ and $\log_{10}\gamma$ for $d = 0.440$, the maximum directivity of the $H_{15,1,e}$ mode equals $D = 8.094$

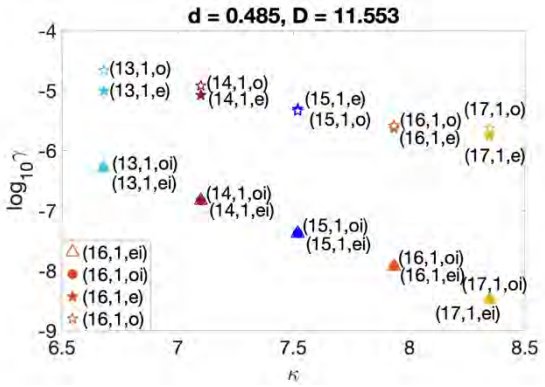


Fig. 5: Values of κ and $\log_{10}\gamma$ for $d = 0.485$, the maximum directivity of the $H_{16,1,e}$ mode equals $D = 11.553$

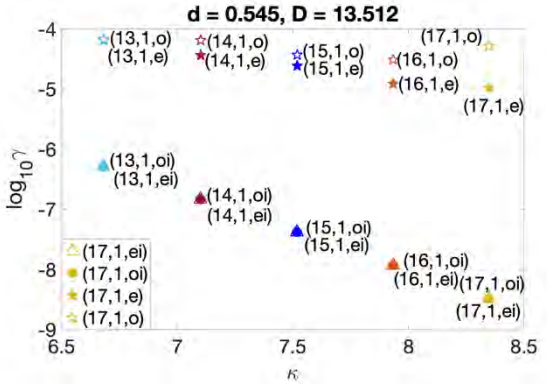


Fig. 6: Values of κ and $\log_{10}\gamma$ for $d = 0.545$, the maximum directivity of the $H_{17,1,e}$ mode equals $D = 13.512$

The analogous results we can see at Fig. 3 - Fig. 6, where solutions for different fixed d corresponding to the maxima of the directivity D for several modes are shown. At Fig. 7 and Fig. 8, we see the dependencies between the position of the hole and the threshold gain, and between the position of the hole and the emission frequency, respectively.

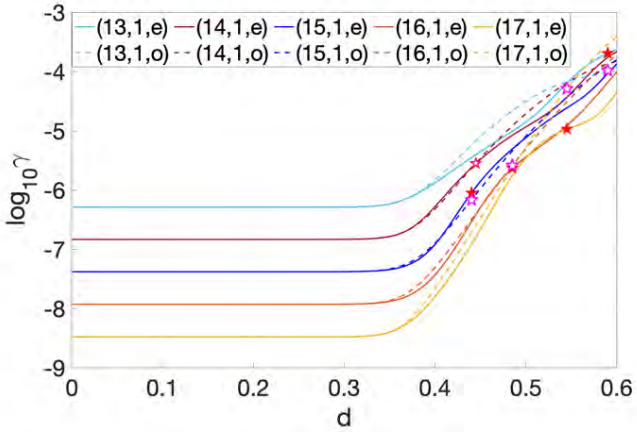


Fig. 7: Dependence of $\log_{10}\gamma$ on d

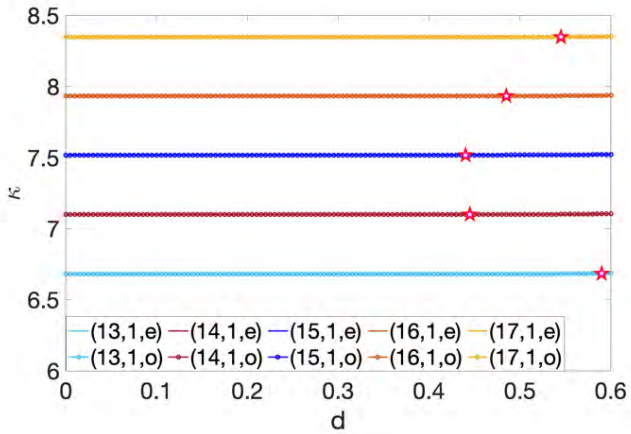


Fig. 8: Dependence of κ on d

At Fig. 9,13,16,19,22, we see the dependencies of the directivity D of the first solution for several modes (even and odd) on d. It is needed to notice that the maxima directivity is really high, thus, it give us a chance to obtain unidirectional emission of lasing. It will be proved by the diagrams of the dependence on d of the angle β of the main beam direction with the maximum directivity shown at Fig. 10,14,17,20,23. At Fig. 9, we see markers that belongs to the first and the second maxima of the directivity. The same notations will be used at other diagrams.

As we can see at Fig. 10,14,17,20,23, a probability of unidirectional emission exists and is quite high. It can be proved by numerical calculations. Let us check this assumption and have a look at the diagrams of far field and near field for the 13th mode (Fig. 11,12).

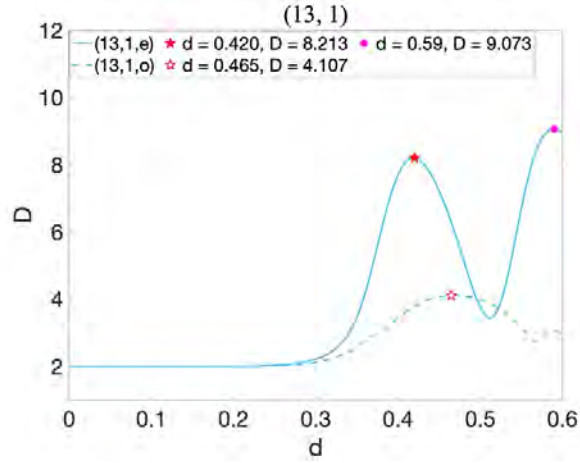


Fig. 9: Dependence of the directivity D of the even and odd $H_{13,1}$ modes on d

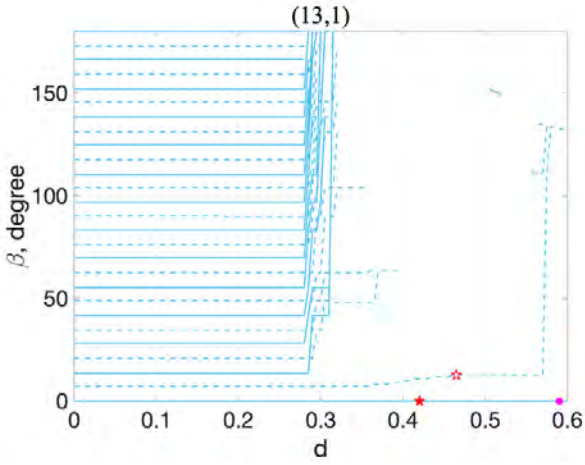


Fig. 10: Dependence of β on d for the even and odd $H_{13,1}$ modes

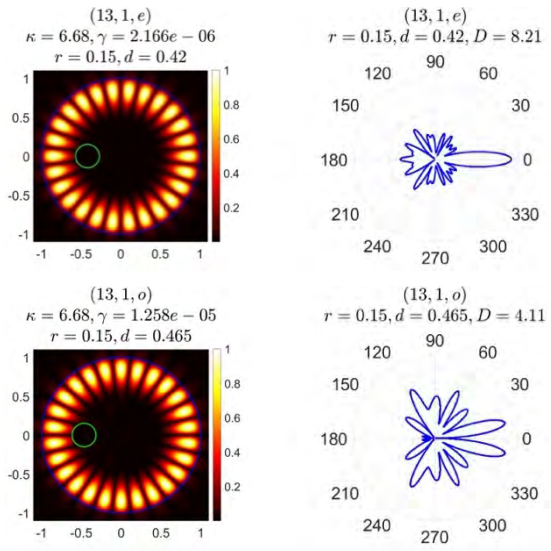


Fig. 11: Diagrams of the fields of the $H_{(13,1,o)}$ and $H_{(13,1,e)}$ modes for $d = 0.465$ and $d = 0.42$

As we can see at Fig. 11, it is impossible to achieve unidirectional emission for odd solution, in this case we will consider to obtain it further only for even solutions. Let us demonstrate analogous results for other higher azimuthal order modes. The results will be shown in comparison of two different maxima .

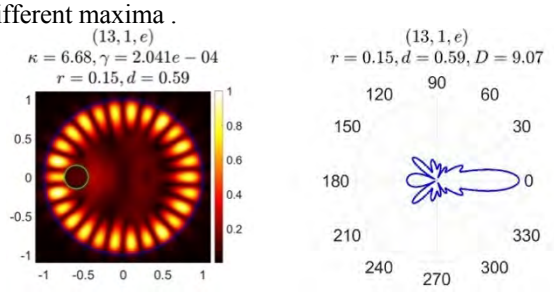


Fig. 12: Diagrams of the fields of the $H_{(13,1,e)}$ mode for $d = 0.59$

At Fig. 13 we see that the maximum directivity was obtained for the point $d = 0.445$.

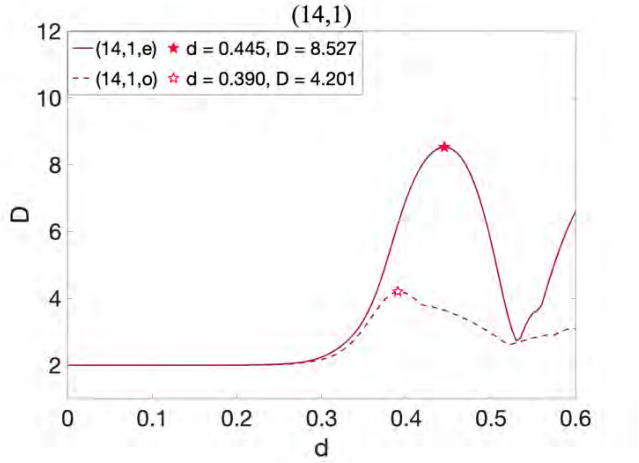


Fig. 13: Dependence of the directivity D of the even and odd $H_{14,1}$ modes on d

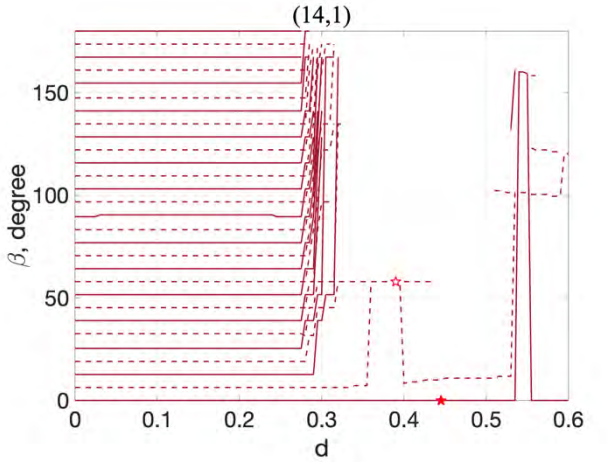


Fig. 14: Dependence of β on d for the even and odd $H_{14,1}$ modes

At Fig. 15, we can also obtain unidirectional lasing emission for even solution of the 14th mode, which corresponds to the results shown at Fig. 14.

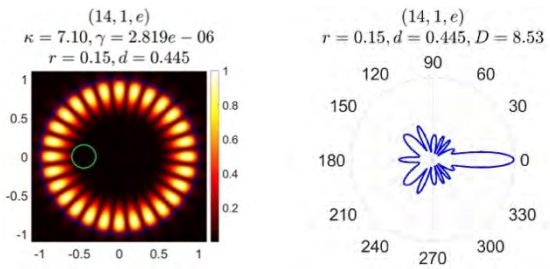


Fig. 15: Diagrams of the fields of the $H_{(14,1,e)}$ mode for $d = 0.445$

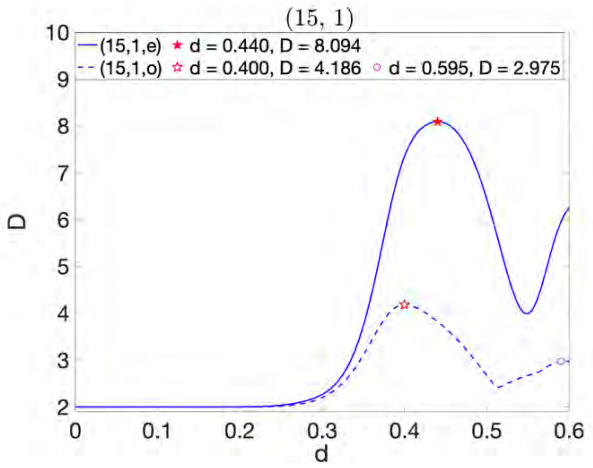


Fig. 16: Dependence of the directivity D of the even and odd $H_{15,1}$ modes on d

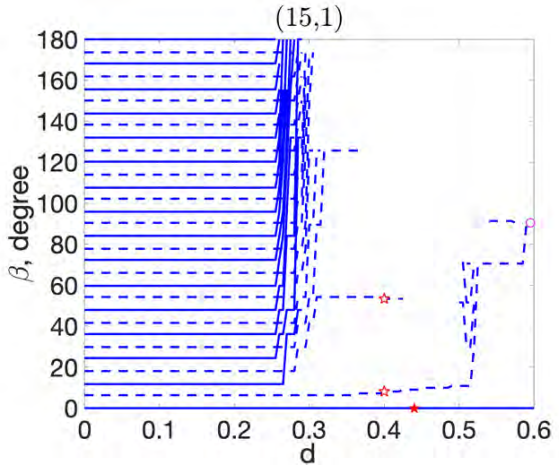


Fig. 17: Dependence of β on d for the even and odd $H_{15.1}$ modes

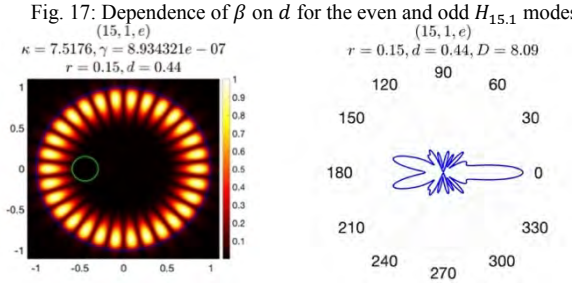


Fig. 18: Diagrams of the fields of the $H_{(15,1,e)}$ mode for $d = 0.44$

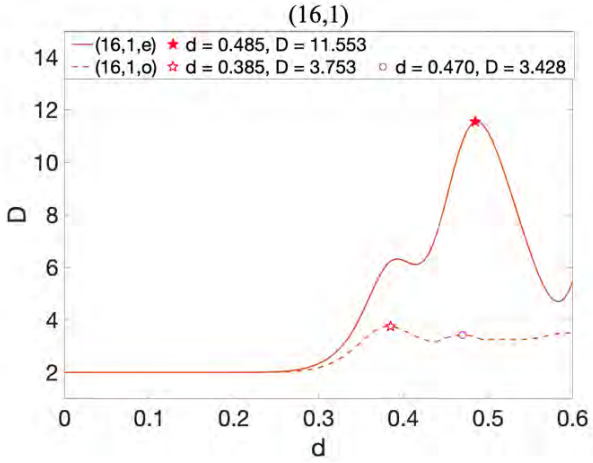


Fig.19: Dependence of the directivity D of the even and odd $H_{16,1}$ modes on d

Looking at Fig. 16,17, we can find that the highest directivity of emission belongs to the first point of maximum. At Fig. 18, the near and far fields for the maximum point are shown.

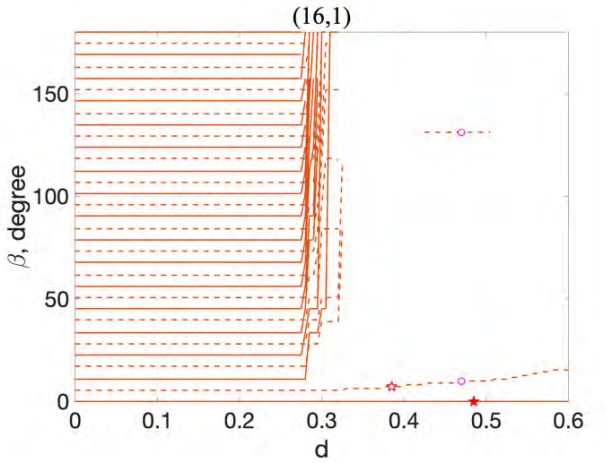


Fig. 20: Dependence of β on d for the even and odd $H_{16,1}$ modes

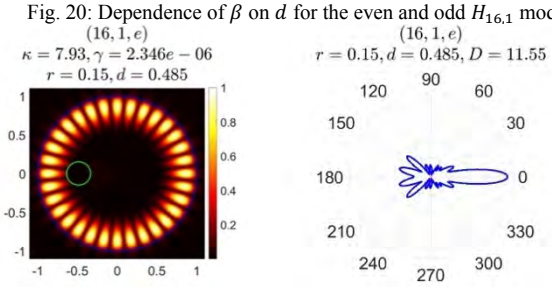


Fig. 21: Diagrams of the fields of the $H_{(16,1,e)}$ mode for $d = 0.485$

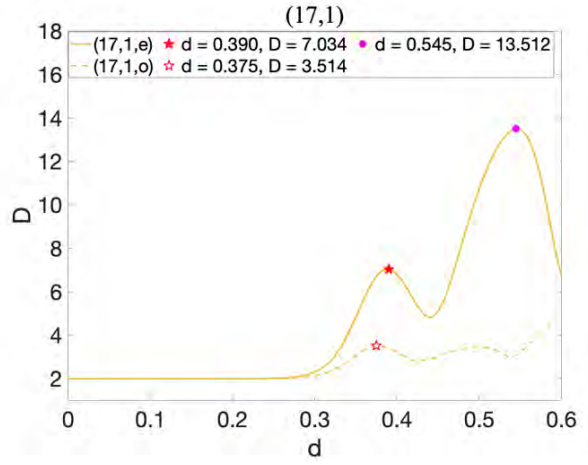


Fig. 22: Dependence of the directivity D of the even and odd $H_{17,1}$ modes on d

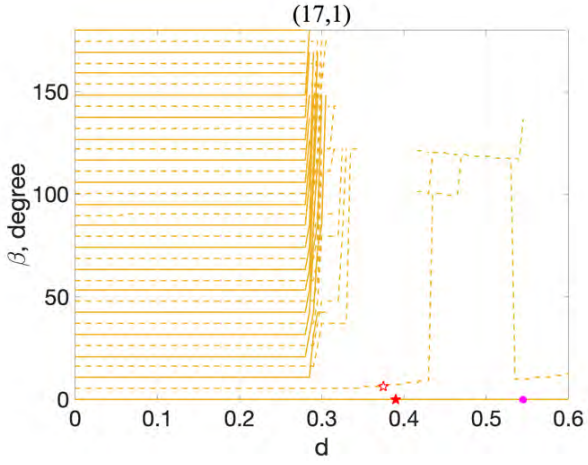


Fig. 23: Dependence of β on d for the even and odd $H_{17,1}$ modes

Looking at Fig. 11,12,15,18,21,24,25, we see that our assumption is confirmed and the maximum directivity corresponds to the unidirectional emission of lasing. At Fig. 22,23, we see the big difference between two maxima, which affects on diagrams of the far field (see Fig. 24,25).

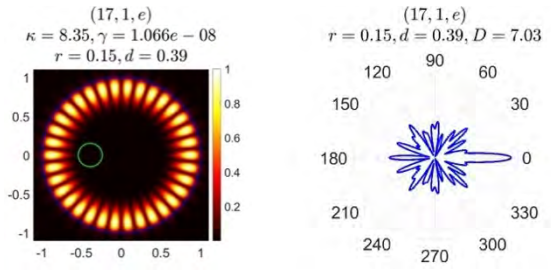


Fig. 24: Diagrams of the fields of the $H_{(17,1,e)}$ mode for $d = 0.39$

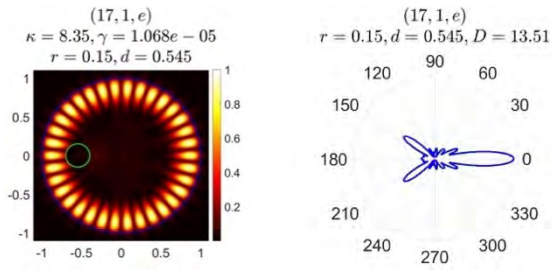


Fig. 25: Diagrams of the fields of the $H_{(17,1,e)}$ mode for $d = 0.545$

IV. Conclusion

Numerical experiments have demonstrated that the unidirectional emission is possible with a certain combination of the hole size and position. Also it is needed to notice that the high directivity corresponds to the low threshold value. In previous works [14,27,28], modes of lower azimuthal orders m were investigated, namely, for $m = 9, 10, 11$. The directivity values D were less than 5.17 and along with one beam in the far field diagram there were several more pronounced beams. This modes can be called quasi-unidirectional. In this work for higher-order modes, higher directivity of emission was obtained and this modes can be called unidirectional, since they have one pronounced ray in the directional pattern.

References

[1] K. J. Vahala, "Optical microcavities," Nature, vol. 424, no. 6950, pp. 839–846, 2003.

[2] L. He, Ş. K. Özdemir, and L. Yang, "Whispering gallery microcavity lasers," Laser and Photonics Reviews, vol. 7, no. 1, pp. 60–82, 2013.

[3] M. Lebental, E. Bogomolny, and J. Zyss, "Organic micro-lasers: a new avenue onto wave chaos physics," in A.B. Matsko (Ed.) Practical Applications of Microresonators in Optics and Photonics, CRC Press, Roca Baton, pp. 317–353, 2009.

[4] J. Wiersig, S. W. Kim and M. Hentschel, "Asymmetric scattering and nonorthogonal mode patterns in optical microspirals," Phys. Rev. A, vol. 78, art. no. 053809, 2008.

[5] E. I. Smotrova, A. I. Nosich, T. M. Benson, and P. Sewell, "Cold-cavity thresholds of microdisks with uniform and non-uniform gain: quasi-3D modeling with accurate 2D analysis," IEEE J. Sel. Topics Quant. Electron., vol. 11, pp. 1135–1142, 2005.

[6] E. I. Smotrova and A. I. Nosich, "Mathematical study of the two-dimensional lasing problem for the whispering-gallery modes in a circular dielectric microcavity," Opt. Quant. Electron, vol. 36, no. 1, pp. 213–221, 2004.

[7] E. I. Smotrova, V. O. Byelobrov, T. M. Benson, J. Ctyroky, R. Sauleau, and A. I. Nosich, "Optical theorem helps understand thresholds of lasing in microcavities with active regions," IEEE J. Quant. Electron., vol. 47, no. 1, pp. 20–30, 2011.

[8] E. I. Smotrova and A. I. Nosich, "Thresholds of lasing and modal patterns of a limacon cavity analysed with Muller's integral equations," Proc. Int. Conf. Laser Fiber-Opt. Networks Modeling (LFNM-2011), Kharkiv, 2011, art. no. 083.

[9] E. I. Smotrova, V. Tsvirkun, I. Gozhyk, C. Lafargue, C. Ulysse, M. Lebental, and A. I. Nosich, "Spectra, thresholds, and modal fields of a kite- shaped microcavity laser," J. Opt. Soc. Am. B, vol. 30, no. 6, pp. 1732–1742, 2013.

[10] A. O. Spiridonov, E. M. Karchevskii, and A. I. Nosich, "Mathematical and numerical modeling of on-threshold modes of 2-D microcavity lasers with piercing holes," Axioms, vol. 8, no. 3, pp. 1–16, 2019.

[11] A. S. Zolotukhina, A. O. Spiridonov, E. M. Karchevskii, and A. I. Nosich, "Electromagnetic analysis of optimal pumping of a microdisc laser with a ring electrode," Appl. Phys. B., vol. 123, no. 1, art. no. 32, 2017.

[12] A. Repina and A. Oktyabrskaya, "Mathematical modeling of photonic crystal resonators based on the Lasing Eigenvalue Problem," Proc. Int. Conf. Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA-2019), Lipetsk, 2019, pp. 472–477.

[13] A. O. Oktyabrskaya, A. O. Spiridonov and E. M. Karchevskii, "Numerical modeling of active microcavities with piercing holes using Muller boundary integral equations and the Galerkin method," Proc. Int. Conf. Days in Diffraction (DD-2019), Saint-Petersburg, 2019, pp. 138–143.

[14] A. Oktyabrskaya, A. Repina and E. Karchevskii, "Laser modes od active circular microcavity with circular piercing hole," Proc. Int. Conf. on electronics and nanotechnology (ELNANO-2020), Kyiv, 2020, pp. 207–210.

[15] A. O. Oktyabrskaya, A. O. Spiridonov and E. M. Karchevskii, "Muller Boundary Integral Equations for Solving Generalized Complex-Frequency Eigenvalue Problem," Lobachevskii J. Math., vol. 41, no. 7, pp. 1377–1384, 2020.

[16] O. V. Shapoval, K. Kobayashi, and A.I.Nosich, "Electromagnetic engineering of a single-mode nanolaser on a metal plasmonic strip placed into a circular quantum wire," IEEE J. Sel. Topics Quant. Electron., vol. 23, no. 6, art. no. 1501609, 2017.

[17] D. M. Natarov, T. M. Benson, and A. I. Nosich, "Electromagnetic analysis of the lasing thresholds of hybrid plasmon modes of a silver tube nanolaser with active core and active shell," Beilstein J. Nanotechnol., vol. 10, pp. 294–304, 2019.

[18] V. O. Byelobrov, J. Ctyroky, T. M. Benson, R. Sauleau, A. Altintas, and A. I. Nosich, "Low-threshold lasing eigenmodes of an infinite periodic chain of quantum wires," Opt. Lett., vol. 35, no. 21, pp. 3634–3636, 2010.

[19] V. O. Byelobrov, T. M. Benson, , and A. I. Nosich, "Binary grating of subwavelength silver and quantum wires as a photonic-plasmonic lasing platform with nanoscale elements," IEEE J. Sel. Top. Quantum Electron, vol. 18, no. 6, pp. 1839–1846, 2012.

[20] S. Zhang Y. Li, P. Hu, A. Li, Y. Zhang, W. Du, M. Du, Q. Li, and F. Yun, "Unidirectional emission of GaN-based eccentric microring laser with low threshold," Opt. Expr., vol. 28, no. 5, pp. 6443–6451, 2020.

[21] A. O. Spiridonov and E. M. Karchevskii, "Mathematical and numerical modeling of a drop-shaped microcavity laser," Comp. Res. Modeling, vol.11, no. 6, pp. 1083–1090, 2019.

[22] A. O. Spiridonov, E. M. Karchevskii, T. M. Benson, and A. I. Nosich, "Why elliptic microcavity lasers emit light on bow-tie-like modes instead of whispering-gallery-like modes," Opt. Comm., vol. 439, pp.112–117, 2019.

[23] A. O. Spiridonov, E. M. Karchevskii, and A. I. Nosich, "Symmetry accounting in the integral-equation analysis of the lasing eigenvalue problem for two-dimensional optical microcavities," J. Opt. Soc. Am. B., vol. 34, pp. 1435–1443, 2017.

[24] A. O. Spiridonov, E. M. Karchevskii, and A. I. Nosich, "Rigorous formulation of the lasing eigenvalue problem as a spectral problem for a Fredholm operator function," Lobachevskii J. Math., vol. 39, no. 8, pp. 1148–1157, 2018.

[25] E. M. Karchevskii, "The fundamental wave problem for cylindrical dielectric waveguides," Differential Equations, vol. 36, no. 7, pp. 1109–1111, 2000.

[26] A. O. Spiridonov and E. M. Karchevskii, "Mathematical and numerical analysis of the spectral characteristics of dielectric microcavities with active regions," Proc. Int. Conf. Days on Diffraction (DD-2016), Saint-Petersburg, 2016, art. no. 7756880, pp. 390-395.

[27] A. I. Repina, A. O. Oktyabrskaya, I. V. Ketov and E. M. Karchevskii, "Laser Modes of Active Eccentric Microring Cavities," Proc. Int. Conf. on Transparent Optical Networks (ICTON-2020), Bari, 2020, art. no. We.C4.6, in press.

[28] A. O. Oktyabrskaya, A. I. Repina, A. O. Spiridonov, E. M. Karchevskii and A. I. Nosich, "Numerical modeling of on-threshold modes of eccentric-ring microcavity lasers using the Muller integral equations and the trigonometric Galerkin method," Opt. Comm., vol. 476, art. no. 126311, 2020.

Implementing a Virtual Network on the SDN Data Plane

Igor Burdonov
Software Engineering department
Ivannikov Institute for System Programming of RAS
Moscow, Russia
igor@ispras.ru

Nina Yevtushenko
Software engineering deparment
Ivannikov Institute for System Programming
Moscow, Russia
evtushenko@ispras.ru

Alexandr Kossachev
Software Engineering department
Ivannikov Institute for System Programming of RAS
Moscow, Russia
kos@ispras.ru

***Abstract*—The paper investigates the implementation of virtual networks on the SDN data plane, modeled by a graph of physical connections between network nodes. A virtual network is defined as a set of ordered host pairs (sender, receiver), and it is implemented by a set of host-host paths that uniquely determine the switch settings. It is shown that any set of host pairs can be implemented on a connected graph without the occurrence of an infinite transfer of packets in a loop and without duplicate paths when the host receives the same packet several times. However, undesired paths may occur when a host receives a packet that is not intended for this host. On the other hand, it is shown that in some cases, implementation without undesired paths inevitably leads to duplication or looping of packets. The question is posed: on which graph can any set of host pairs be implemented without looping, duplication and undesired paths? A sufficient condition is proposed and a hypothesis is put that this condition is also the necessary condition.**

Keywords— Software Defined Networking (SDN), data plane, Network connectivity topology

I. Introduction

Software Defined Networking (SDN) is one of the main technologies for network virtualization [1][2][3][4][5]. On the data plane, packets are transmitted between hosts through intermediate switches. This is modeled by a graph of physical connections (links) often referred to as RNCT (Resource network connectivity topology), the vertices of which are hosts and switches, and the edges correspond to physical connections between them. The switches are configured by SDN controller (-s), setting up a set of flow rules for each switch. The rule determines which neighboring vertices of the graph the packet received by the switch is forwarded to, depending on which neighbor the packet came from and the parameter vector in the packet header [6]. Thus, the configuration of the network switches determines the set of paths from host to host, through which packets will be forwarded.

There are tasks of two levels. 1) How to implement a given set of host-host paths through appropriate switch settings? 2) How to implement a given set of pairs (host, host) through appropriate host-host paths in the graph of physical connections?

It is known that when solving the 1st problem there are three effects. 1a) Cycles may occur in which packets will be transmitted endlessly and, moreover, endlessly cloned. 1b) Undesired paths may appear. These problems were investigated in [4] [5]. 1c) Duplicate paths may appear, due to which the host destination receives the same packet more than once.

The 2nd task is reduced to the 1st task by choosing a suitable set of paths, avoiding the above effects if possible. The following questions are then raised. 2a) Is it possible to implement a given set of host pairs on a graph, i.e. to choose a suitable set of paths without the indicated effects? 2b) Is it possible to implement any set of host pairs without such effects on a given graph of physical connections?

As an answer to question 2a), this paper shows that any set of host pairs can be implemented on a connected graph without cycles and duplication, however, undesired paths may occur, i.e. paths connecting unintended host pairs. On the other hand, the article demonstrates that in some cases, the implementation of a given set of host pairs without undesired paths inevitably leads to path duplication or loops. The article establishes a sufficient condition for the positive answer to question 2b) and puts the hypothesis that this sufficient condition is also the necessary condition.

II. Preliminaries

A physical connection graph (hereinafter simply a graph), often referred to as RNCT (Resource network connectivity topology), is a connected undirected graph $G = \{V, E\}$ without multiple edges and loops, where V is the set of switches and hosts, $E \subseteq V \times V$ is the set of edges modeling physical connections between the switches and hosts. Since the edge connecting the vertices a and b is undirected and there are no multiple edges, it can be denoted by both ab and ba. Since there are no loops, there are no edges of the form aa in E. Since there are no multiple edges, a path as a sequence of adjacent edges is uniquely determined by the sequence of vertices $a_1 \ldots a_n$ through which it passes. A path starting at vertex a and ending at vertex b is called an ab-path. If the path passes along the edge ab from a to b, then we say that it passes through the arc ab. If a and b are hosts, an ab-path in which all vertices except the first vertex a and the last vertex b are switches, is called a *complete path*. A

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

path is called *vertex-simple* (*edge-simple*) if each vertex (arc) occurs at most once. The vertices of the graph will be denoted by lowercase letters a, b, c,…x, y, z, the paths by bold lowercase letters $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$,…, and the set of paths by capital letters P, Q, R, …

We will assume that each host x is connected to exactly one switch [3]. Therefore, the host is the terminal vertex of the graph, i.e. a vertex of degree 1. If the switch a has degree 1 and is connected to vertex b, then any complete path passing through a has the form ...*bab*...; removing all *bab* cycles from it, we get a path that does not pass through a. This means that such a switch is "superfluous", and it is enough to consider graphs in which terminal vertices are only hosts. The sets of hosts and switches are denoted by H and S, respectively; $H \cup S = V$, $H \cap S = \varnothing$.

In general case, the rule of the switch b has the form σabc, where a and c are neighbors of b, and σ is the vector of packet header parameters that can be used in the rules. Such a rule means that switch b, having received a packet with vector σ from neighbor a, forwards it to neighbor c. It is assumed that the switch does not change σ. Thus, for the vector σ complete paths of the form $a_1 \ldots a_n$ are considered where in the switch a_i there is a rule $\sigma a_{i-1} a_i a_{i+1}$, $i = 2..n-1$. If there are two rules σabc and $\sigma abc`$, where $c \neq c`$, then it is said that the packet is cloned, i.e. is sent to both neighbors c and $c`$.

The given set P of complete paths uniquely determines the minimal set of switch rules that induces all paths from P. However, this does not mean that only paths of P are induced. We say that two paths are *merging* paths on the arc ab at vertex a if they have an intermediate common arc ab with different direct predecessor arcs ca and $c`a$ where $c \neq c`$, and are *separating paths* after the arc de at the vertex e if they have an intermediate common arc de with different direct successor arcs eg and $eg`$ where $g \neq g`$.

There is a cycle in the path if a complete xy-path passes through an arc twice, i.e. the path has the form $\boldsymbol{paqer}(\boldsymbol{aqer})^{*}\boldsymbol{aqes}$, where the segment $\boldsymbol{p}$ starts at the host $x \neq a$, the segments $\boldsymbol{p}$ and $\boldsymbol{r}$ do not end at the same vertex, after these segments the switch a follows, the segment $\boldsymbol{aqer}$ passes one or more times, after the switch e segments $\boldsymbol{r}$ and $\boldsymbol{s}$ do not start at one vertex, and segment $\boldsymbol{s}$ ends at the host $y \neq e$. Moving along the path, we see that at the vertex a the path merges with itself, then at the vertex e it separates with itself, and then this merging and separating occurs again. If the path goes through the cycle k times, then it will be $k + 1$ times both separating after merging, and merging after separating. Packets will not only endlessly traverse the $\boldsymbol{aqer}$ cycle, but also endlessly clone at vertex e, so host y will receive an infinite number of clones of the same packet.

A path that does not merge with itself is an edge-simple path. For the absence of cycles, it is necessary that all paths of the set P be edge-simple. But that is not sufficient. If two edge-simple complete paths from P after merging on the arc ab are separated (after this or another arc), i.e. those are $x\boldsymbol{p}ab\boldsymbol{q}y$ and $x`\boldsymbol{p}`ab\boldsymbol{q}`y`$ with different start and end hosts $x \neq x`$ and $y \neq y`$, then new paths $x\boldsymbol{p}ab\boldsymbol{q}`y`$ and $x`\boldsymbol{p}`ab\boldsymbol{q}y$ are also induced. This operation of inducing new paths is called the *arc closure*, and the result of the arc closure of all pairs of paths from P is denoted by $P{\downarrow}{\uparrow}$ [4] [5]. Obviously, $P \subseteq P{\downarrow}{\uparrow}$. If $P \neq P{\downarrow}{\uparrow}$, i.e. P is not arc closed, then undesired paths occur. In particular, non-edge-simple paths and, therefore, cycles may occur. The appearance of cycles in the arc closure of the set of complete paths always indicates the infinity of this arc closure and, thus, the presence of duplication. There are no cycles in a finite arc closed set of complete edge-simple paths.

For a set of complete paths P, by $H(P) \subseteq H \times H$ we denote the set of pairs xy for which there is an xy-path in P. A set of host pairs $D \subseteq H \times H$ that does not contain pairs of the form xx will be called *normal*. We say that a normal set D (*non-strictly*) is implemented by an arc closed set of complete paths P if $D \subseteq H(P)$, D is *strictly* implemented if $D = H(P)$, D is implemented *without cycles* if P is finite, D is implemented *without duplication* if P contains exactly one xy-path for each pair $xy \in D$.

If the source address is included into the parameter vector σ, then the rules for parameter vectors with different source addresses work independently. For each source host x in the graph, the tree I_x of shortest paths leading from x to all other hosts can be selected. For any normal set D of host pairs and any host x, a subset D_x is selected, where the first element of the pair is host x, and the subtree $I_x(D)$ is selected, in which leaf vertices are destination hosts y such that $xy \in D_x$. In the outgoing tree, all paths are edge-simple (even vertex-simple, that is, without vertex repetition), and there is no merging, thereby there is no separating after merging. Therefore, $I_x(D)$ is arc closed and, obviously, strictly implements D_x without cycles and duplication; moreover, the shortest complete paths are used. Thus, in this case there is no problem with the strict implementation without loops and duplication of any normal set of host pairs. Moreover, the implementation of any such set turns out to be a subset of the same set of paths, namely, the union of I_x trees over all source hosts x. A similar procedure with a similar result is applicable when the destination address is included into the parameter vector σ. In this case, the incoming O_x tree is built for each destination host x.

Below we consider the case when the source address and the destination address are not included into the parameter vector σ. The remaining parameters do not affect packet transmission with the given vector σ, so we will omit σ in the designation of the rule and write abc instead of σabc. In other words, the switch rules (for a given vector σ) determine to which vertex the packet should be sent, only depending on the neighbor from which the packet was received. In this case, the maximum number of rules by which the switch operates depends only on the number of its neighbors and does not depend on the number of hosts in the network.

III. Non-strict / strict implementation of versus cycles and duplication

In this section, we examine the relationship between the non-strict and strict implementation of the set of host pairs with the presence or absence of cycles and duplication.

Proposition 1. On a connected graph G, any normal set D of host pairs is non-strictly implemented without cycles and duplication.

Proof. In G, choose an arbitrary spanning tree T. Since a host has degree 1 in G, all the hosts are leaves of T. Let P be the set of all shortest complete paths in the tree T. Obviously, all paths from P are vertex-simple and, therefore, edge-simple; moreover, there are no duplicate paths and P is finite and arc closed. If we leave only xy-paths in the set P such that $xy \in D$, then for the resulting set $P(D)$ we have $P(D)\downarrow\uparrow \subseteq P$. Thus, in $P(D)\downarrow\uparrow$ all the paths are also edge-simple, there are no duplicate paths and $P(D)\downarrow\uparrow$ is finite and arc closed. By construction, $D = H(P(D)) \subseteq H(P(D)\downarrow\uparrow)$. Therefore, $P(D)\downarrow\uparrow$ non-strictly implements D without cycles and duplication.

For a complete path $\boldsymbol{p}$, let $\boldsymbol{p}^\circ$ denote the path that is obtained from $\boldsymbol{p}$ by using, as far as possible, the following operation to delete cycles: the path $\boldsymbol{p} = \boldsymbol{qaras}$ turns into the path $\boldsymbol{p}^\circ = \boldsymbol{qas}$. Note that the result of the operation "°", generally speaking, is ambiguous. For the sake of simplicity, we assume that the operation of deleting a cycle is applied only when each vertex of the prefix $\boldsymbol{q}$ occurs only once in $\boldsymbol{p}$ and $\boldsymbol{q}$ and $\boldsymbol{r}$ do not contain vertex a. In other words, if the vertex a occurs first among the vertices that have several occurrences in $\boldsymbol{p}$, then a cycle is removed by deleting $\boldsymbol{ra}$ in $\boldsymbol{p}$, where $\boldsymbol{r}$ does not contain $\boldsymbol{a}$. For example, for $\boldsymbol{p} = xacbcaby$, $\boldsymbol{p}^\circ = xaby$ (not $xacby$) is obtained. The procedure ° terminates when each vertex occurs at most once in $\boldsymbol{p}^\circ$. Given a set of complete paths P, P° is the set of paths obtained by deleting cycles from all paths P, i.e. $P^\circ = \{ \boldsymbol{p}^\circ \mid \boldsymbol{p} \in P \}$.

Proposition 2. Let P be the set of complete paths. Then P° consists of vertex-simple paths and connects the same pairs of hosts that the set P: $H(P^\circ) = H(P)$. If P is arc closed, then the arc closure $P^\circ\downarrow\uparrow$ connects the same pairs of hosts: $H(P^\circ\downarrow\uparrow) = H(P^\circ)$, i.e. the arc closure $P^\circ\downarrow\uparrow$ adds only duplicate paths to the set P°. If P is finite and arc closed, then P° is finite and arc closed.

Proof. Obviously, removing all cycles from a path makes the path vertex-simple. Therefore, P° consists of vertex-simple paths. The operation of deleting one cycle from one path does not change $H(P)$. Therefore, the chain of such operations also does not change $H(P)$. Therefore, P° connects the same host pairs that the set P: $H(P^\circ) = H(P)$.

Let us prove that if the set P is arc closed, then $H(P^\circ\downarrow\uparrow) = H(P^\circ)$. Indeed, let the set P° contain xy-path $\boldsymbol{p}$ and $x`y`$-path $\boldsymbol{q}$, which are obtained by removing cycles from xy-path $\boldsymbol{p}`$ and $x`y`$-path $\boldsymbol{q}`$, respectively, which are elements of the set P. If the path $\boldsymbol{p}$ and $\boldsymbol{q}$ have a common arc, then this arc is also common for the paths $\boldsymbol{p}`$ and $\boldsymbol{q}`$. Since P is arc closed, it also contains an $xy`$-path and an $x`y$-path. Thus, after deleting the cycles in P° there will also be an $xy`$-path and an $x`y$-path. Therefore, $H(P^\circ\downarrow\uparrow) = H(P`)$.

Let P be finite and arc closed. Then, obviously, P° is also finite. Let us prove that P° is arc closed. Let P° contain paths $\boldsymbol{pabq}$ and $\boldsymbol{p}_1\boldsymbol{abq}_1$ with a common arc. These paths are obtained by deleting the cycles from the paths $\boldsymbol{r}$ and $\boldsymbol{r}_1$, respectively, which are elements of the set P. Since, when deleting the cycles, any nonempty sequence between any two occurrences of the vertices a and b can be completely deleted only together with the removal of the occurrence a and/or b, the paths $\boldsymbol{r}$ and $\boldsymbol{r}_1$ can be represented as $\boldsymbol{p}`\boldsymbol{abq}`$ and $\boldsymbol{p}`_1\boldsymbol{abq}`_1$, respectively, where $\boldsymbol{p} = \boldsymbol{p}`^\circ$, $\boldsymbol{p}_1 = \boldsymbol{p}`_1{}^\circ$, $\boldsymbol{q} = \boldsymbol{q}`^\circ$, $\boldsymbol{q}_1 = \boldsymbol{q}`_1{}^\circ$. Since P is arc closed, P contains the paths $\boldsymbol{p}`\boldsymbol{abq}_1`$ and $\boldsymbol{p}`_1\boldsymbol{abq}`$. And then P° contains the paths $\boldsymbol{pabq}_1 = \boldsymbol{p}`^\circ\boldsymbol{abq}_1`^\circ = (\boldsymbol{p}`\boldsymbol{abq}_1`)^\circ$ and $\boldsymbol{p}_1\boldsymbol{abq} = \boldsymbol{p}`_1{}^\circ\boldsymbol{abq}`^\circ = (\boldsymbol{p}`_1\boldsymbol{abq}`)^\circ$. Therefore, the set P° is arc closed.

Note that all conditions on the set P of complete paths in Proposition 2 are necessary conditions. If the set P is not arc closed, then the arc closure $P^\circ\downarrow\uparrow$ can connect additional pairs of hosts: $H(P^\circ\downarrow\uparrow) \supseteq H(P^\circ)$, i.e. arc closure $P^\circ\downarrow\uparrow$ can add to the set P° not only duplicate paths. If the set P is finite, but not arc closed, then the set P° is finite, but not necessary is arc closed. In both cases, the set $P = \{ xaby, x`aby` \}$ can serve as an example, where $x \neq x`$ and $y \neq y`$: $P^\circ = P$, $P^\circ\downarrow\uparrow = \{ xaby, x`aby`, xaby`, x`aby \}$. If the set P is arc closed, but infinite, then the set P° not necessary is arc closed. As an example the set $P = Q\downarrow\uparrow$ can be considered where $Q = \{ xabcdy, x`cdaby` \}$, $x \neq x`$ and $y \neq y`$: the set P contains the path $xabcdaby`$ that is not edge-simple, $P^\circ = Q$ and $P^\circ\downarrow\uparrow = P$.

It follows from Proposition 2 that any set D of host pairs that is strictly implemented without cycles can be strictly implemented by a set of vertex-simple paths. It is sufficient to take the set P° of vertex-simple paths instead of a finite arc closed set P of complete paths that strictly implements D.

Proposition 3. A strict implementation is not always possible without duplication: there is a graph on which some normal set of host pairs is strictly implemented only with duplication.

Proof. Consider the example in Figure 1. The set D is strictly implemented by a finite arc closed set of paths P that contains duplicate paths $x_0a_0a_1b_1b_0y_0$ and $x_0a_0a_2b_2b_0y_0$. Let us assume that the arc closed set of paths $P`$ without duplicate paths strictly implements the set $D$. Since $D$ contains 7 pairs of hosts, $P`$ strictly implements D and there are no duplicate paths in $P`$, $P`$ must contain exactly 7 paths. Then, by Proposition 2, we can choose the set $P`$ consisting only vertex-simple paths. In order to reach host y_j from host x_i, the path must go either via the arc a_1b_1 or via the arc a_2b_2. Let m_i paths, $i = 1, 2$, pass via the arc a_ib_i, and these paths start at n_i hosts and end at k_i hosts. Then $n_1 + n_2 = 3$, $k_1 + k_2 = 3$, $m_1 + m_2 = 7$. The set $P`$ is arc closed, which implies $n_1k_1 = m_1$, $n_2k_2 = m_2$. Hence $n_1k_1 + (3 - n_1)(3 - k_1) = 7$, which implies $2n_1k_1 = 3(n_1 + k_1) - 2$. In this example, $k_1 \leq 3$ and $n_1 \leq 3$ which implies: $3k_1 = 2$ for $n_1 = 0$, $k_1 = -1$ for $n_1 = 1$, $k_1 = 4$ for $n_1 = 2$, $3k_1 = 7$ for $n_1 = 3$. Each of these equations has no solution for non-negative integers or contradicts the condition $k_1 \leq 3$. We came to a contradiction, therefore, our assumption is not true, and the proposition is proved.

Proposition 4. A strict implementation is not always possible without cycles: there is a graph on which some normal set of host pairs is strictly implemented, but only by an infinite arc closed path set.

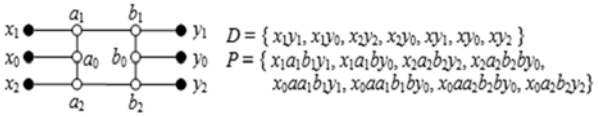


Fig. 1. The set D is strictly implemented only with duplication.

Proof. Consider the example in Figure 2. The set D is strictly implemented by arc closure $P\downarrow\uparrow$ of the set of paths P. But in P there are paths $x_1a_1b_1c_1c_2a_2b_2y_2$ and $x_2a_2b_2d_1d_2a_1b_1y_1$, which in $P\downarrow\uparrow$ induce a non-edge-simple path $x_1a_1b_1c_1c_2a_2b_2d_1d_2a_1b_1y_1$ that goes twice via the arc a_1b_1, and therefore $P\downarrow\uparrow$ is infinite. Let the arc closed set of paths $P`$ strictly implement the set $D$ and $P`$ is finite. Then, by Proposition 2, we can choose the set $P`$ consisting of vertex-simple paths.

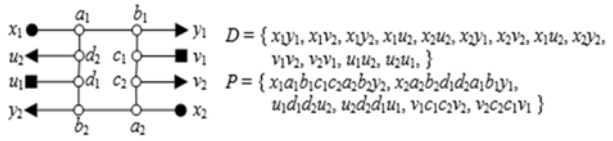


Fig. 2. The set D is strictly implemented only with cycles.

1. Let there be x_2y_1-path $\boldsymbol{p_1}$ passing via the arc c_2c_1.

1.1. Let there be a v_2v_1-path $\boldsymbol{p_2}$ passing via the arc c_2c_1. Then x_2y_1-path $\boldsymbol{p_1}$ and v_2v_1-path $\boldsymbol{p_2}$ both pass via the arc c_2c_1 and, therefore, induce an x_2v_1-path but $x_2v_1 \notin D$.

1.2. Therefore, any v_2v_1-path $\boldsymbol{p_3}$ does not pass via the arc c_2c_1, and then it passes via the arcs a_2b_2 and a_1b_1.

1.2.1. Let there be an x_1y_1-path $\boldsymbol{p_4}$ passing via the arc a_1b_1. Then v_2v_1-path $\boldsymbol{p_3}$ and x_1y_1-path $\boldsymbol{p_4}$ both pass via the arc a_1b_1 and, therefore, induce a v_2y_1-path but $v_2y_1 \notin D$.

1.2.2. Therefore, any x_1y_1-path $\boldsymbol{p_5}$ does not pass via the arc a_1b_1, and then it passes via the arc c_2c_1.

1.2.2.1. Let there be an x_2y_2-path $\boldsymbol{p_6}$ passing via the arc a_2b_2. Then v_2v_1-path $\boldsymbol{p_3}$ and x_2y_2-path $\boldsymbol{p_6}$ both pass via the arc a_2b_2 and, therefore, induce an x_2v_1-path but $x_2v_1 \notin D$.

1.2.2.2. Therefore, any x_2y_2-path does not pass via the arc a_2b_2. Such a path is unique among vertex-simple paths: $\boldsymbol{p_7} = x_2a_2c_2c_1b_1a_1d_2d_1b_2y_2$. Similarly, any x_1y_1-path does not go via the arc a_1b_1. Such a path is unique among vertex-simple paths: $\boldsymbol{p_5} = x_1a_1d_2d_1b_2a_2c_2c_1b_1y_1$. The paths $\boldsymbol{p_7}$ and $\boldsymbol{p_5}$ have a common arc a_2c_2, therefore the path $x_1a_1d_2d_1b_2a_2c_2c_1b_1a_1d_2d_1b_2y_2$ is induced, which passes twice through the arc a_1d_2, i.e. this path is not edge-simple.

2. Thus, any x_2y_1-path does not pass via the arc c_2c_1. Such a path is unique among vertex-simple paths: $\boldsymbol{p_8} = x_2a_2b_2d_1d_2a_1b_1y_1$. Due to symmetry, it is similarly proved that any x_1y_2-path does not pass via the arc d_2d_1. Such a path is unique among vertex-simple paths: $\boldsymbol{p_9} = x_1a_1b_1c_1c_2a_2b_2y_2$.

The paths $\boldsymbol{p_8}$ and $\boldsymbol{p_9}$ have a common arc a_1b_1, so the path $\boldsymbol{p}_{10} = x_2a_2b_2d_1d_2a_1b_1c_1c_2a_2b_2y_2$ is induced. This path passes twice through the arc a_2b_2, i.e. this path is not edge-simple. We came to a contradiction and, therefore, our assumption is not true, and the proposition is proved.

IV. A SUFFICIENT CONDITION FOR THE STRICT IMPLEMENTATION OF ANY SET OF HOST PAIRS WITHOUT CYCLES AND DUPLICATION

In this section, we investigate sufficient conditions on a graph that allow us to strictly implement any set of host pairs without cycles and duplication. If for two paths there is a merge on the arc ab and there is a separation after the arc cd, then we say that the separation occurs after the merge, if at least one of these paths first passes the arc ab and then the arc cd. Accordingly, merging occurs after separation, if at least one of these paths first passes the arc cd and then the arc ab. Note that in the linear order of the vertices of one path, the separation after the arc cd can occur after merging on the arc ab, and in the linear order of the vertices of the other path, on the contrary, the merging on the arc ab can occur after the separation after the arc cd, as demonstrated by the example of two paths: *xabefcdy* and *ufcdabev*, where different letters indicate different vertices.

Proposition 5. Given a finite set of complete paths, if there is no separation after merging, then there are no cycles but the converse is not always true.

Proof. The sufficiency follows from the fact that the arc closure can induce new paths only in the case of the separation after merging. Also, the cycle is induced by a non-edge-simple path, in which there is a separation after the merging, as indicated in Section 2. But the presence of the separation after merging does not necessarily mean the arc non-closure or the presence of cycles, as demonstrated by the following example of a finite arc closed set of complete paths without cycles $P = \{ xaby, xaby`, x`aby, x`aby` \}$, where the hosts x, y, $x`$ and $y`$ are pairwise different.

Proposition 6. Given an arc closed set P of complete paths, the absence of merging after separation is the necessary and sufficient condition for the absence of duplication.

Proof. Let there be two different xy-paths. Since each host is connected to exactly one switch, the maximum common prefix of these paths and the maximum common postfix of these paths each has length at least 1, and the prefix can be represented as $\boldsymbol{xpa}$, and the postfix as $\boldsymbol{bqy}$. Then, obviously, a is traversed in each of the paths earlier than b, and the paths are of the form $\boldsymbol{xpar bqy}$ and $\boldsymbol{xpar`bqy}$. Thus, these paths separate at the vertex $a$, and then merge at the vertex $b$. From this follows the sufficiency of the condition. Let us prove the necessity of the condition. Let there be a merge after the separation for two complete paths: $xy$-path and $x`y`$-path, with the xy-path first passing through the vertex a at which the paths are separated, and then the vertex b where the paths merge. Two cases are possible.

1) Figure 3 (a). The $x`y`$-path goes through the vertices a and b in the same order: ab. Then the paths have the form $\boldsymbol{xpaqbry}$ and $\boldsymbol{x`p`aq`br`y`}$, where the segments $\boldsymbol{xp}$ and $\boldsymbol{x`p`}$ end at the

same vertex *c*, the segments ***q*** and ***q***` start and end at different vertices, the segments ***r****y* and ***r***`*y*` start at the same vertex *d*. The arc *ca* is common for these paths; therefore, the path *x****paq***`*b****r***`*y*` is in the arc closure. Compare this path with the path *x****paqbr****y*. The arc *db* is common for these paths, so the path *x****paq***`*br****y*** is in the arc closure. Since the segments ***q*** and ***q***` start and end at different vertices, the paths *x****paqbr****y* and *x****paq***`*br****y*** are different, therefore, these are duplicate paths.

2) Figure 3 (b). The *x*`*y*`-path goes through the vertices *a* and *b* in the reverse order: *ba*. Then the paths have the form *x****paqbr****y* and *x*`***p***`*b****q***`*a****r***`*y*`, where the segments *x****p*** and *b****q***` end at the same vertex *c*, the segments ***q*** and ***r***`*y*` start at different vertices, the segments ***q*** and *x*`***p***` end at different vertices, the segments ***r****y* and ***q***`*a* start at the same vertex *d*. The arc *bd* is common for these paths, so the path *x****paqbq***`*a****r***`*y*` is in the arc closure. Compare this path with the path *x****paqbr****y*. The arc *ca* is common for these paths, so the path *x****paqbq***`*aqbry* is in the arc closure. Compare this path with the *x****paqbr****y* path. Since the segment ***q***`*aqb* is not empty, the paths *x****paqbr****y* and *x****paqbq***`*aqbry* are different, therefore, these are duplicate paths.

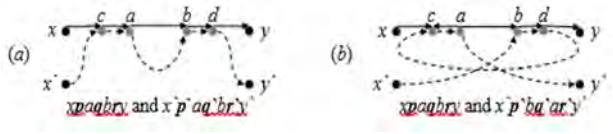


Fig. 3. Merging after separation.

A graph in which any normal set of host pairs can be strictly implemented without cycles is called *almost good*. A graph in which any normal set of host pairs can be strictly implemented without cycles and without duplication is called *good*.

A graph in which any normal set of host pairs can be strictly implemented without cycles is called *almost good*. A graph in which any normal set of host pairs can be strictly implemented without cycles and without duplication is called *good*.

A finite arc closed set of paths *P* connecting all pairs of different hosts (i.e., *H*(*P*) is the largest normal set of host pairs) will be called *almost perfect* if there is no path separation after the path merging, and *perfect* if, in addition, there is no path merging after the path separation. A graph will be called *almost perfect* or *perfect* if it contains, respectively, an almost perfect or perfect set of paths.

A sufficient condition for the strict implementation of any normal set of host pairs without cycles and duplication can now be formulated as the following proposition.

Proposition 7. An almost perfect graph is almost good, and a perfect graph is good. Moreover, the almost perfect set of paths for each normal set of host pairs contains its strict implementation without cycles as a subset, and the perfect set of paths for each normal set of host pairs contains its strict implementation without cycles and without duplication as a subset.

Proof. Since the almost perfect set of paths is finite and there is no separation after merging, any subset of it is also finite and there is no separation after merging, therefore, by Proposition 5, it is arc closed and does not generate cycles. By definition, a perfect set is almost perfect, so any subset of it is also finite, arc closed and does not generate cycles. Since there are no merging after separation in a perfect set of paths, there is no merging after any separation in any of its subsets, and, according to Proposition 6, it does not generate duplication. Given a normal set *D* of host pairs and an almost perfect (perfect) set *P* of paths, we can choose a subset *P*(*D*) such that *H*(*P*(*D*)) = *D*. The set *P*(*D*) of paths strictly implements the set *D* of host pairs without cycles and without duplication if the set *P* is perfect.

V. Conclusions

The paper shows that any set of host pairs can be implemented on a connected graph using paths connecting the hosts of given pairs, without the occurrence of cycles through which packets will circulate endlessly and endlessly multiply, and without duplicate paths, i.e. different paths connecting the same host pairs. However, this may result in extra paths connecting additional host pairs that are not in the given set of pairs. If the absence of undesired paths is required, then for some graphs some sets of host pairs are implemented only with duplication or cycles. The requirements on the graph are formulated and proved, which are sufficient to implement any set of host pairs without cycles (possibly with duplication), and without cycles and without duplication. At the same time, the very possibility of implementing on a graph any set of host pairs without cycles and, especially, without cycles and without duplication seems to be a fairly strong requirement. Therefore, we can hypothesize that these requirements on the graph are also necessary. Confirmation or refutation of this hypothesis is one of the areas for further research.

Acknowledgment

This work is partly supported by RFBR project N 20-07-00338 A.

References

[1] Sezer S., Scott-Hayward S., Chouhan P. K., Fraser B., Lake D., Finnegan J., Viljoen N., Miller M. and Rao N. Are we ready for sdn? Implementation challenges for software-defined networks *IEEE Communications Magazine*, 2013, 51, 7: pp. 36-43.

[2] López J., Kushik N., Yevtushenko N. and Zeghlache D. Analyzing and Validating Virtual Network Requests. *Proc. ICSOFT2017*: pp. 441-6.

[3] Yevtushenko N., Burdonov I., Kossatchev A., Lopez J., Kushik N. and Zeghlache D. Test Derivation for the Software Defined Networking Platforms: Novel Fault Models and Test Completeness *Proc. IEEE East-West Design and Test Symposium*, EWDTS2018, N 8524712: pp. 1-5.

[4] Burdonov I. B., Yevtushenko N. V. and Kossatchev A. S. Testing switch rules in softwaredefined networks., *Trudy ISP RAN/Proc. ISP RAS*, 2018, vol. 30, issue 6: pp. 69-88 (in Russian).

[5] Burdonov I., Kossachev A., Yevtushenko N., López J., Kushik N. and Zeghlache D. Verifying SDN Data Path Requests, 2019, *CoRR abs/1906.03101*.

[6] Boufkhad Y., De La Paz R., Linguaglossa L., Mathieu F, Perino D. and Viennot L. Forwarding tables verification through representative header sets, 2016, *arXiv preprint arXiv:1601.07002*.

Determining the Direction of True Meridian by Micromechanical Gyro

Vladimir Bogolyubov
Kazan National Research Technical University
Kazan, Russia
bvm200@yandex.ru

Lyalya Bakhtieva
Kazan Federal University
Kazan, Russia
lbakhtie@yandex.ru

***Abstract*—The gyrocompassing method based on the parametric excitation of micromechanical gyro (MMG) is proposed. The incoherent mode of parametric excitation of MMG mounted on a horizontal base is considered. A feature of this mode is the presence of "strong resonance", which enhances the amplitude of oscillations of the rotor with respect to its resonance value, and "weak resonance", which reduces its amplitude of oscillations. The vibrational shape of the rotor is not preserved in this mode, and instead of harmonic vibration with one spectral component, there is a complex vibration (observed as a quasi-harmonic vibration) consisting of two spectral components with a slight difference in their frequencies. Thus, the rotor vibrations have the shape and frequency of the generated oscillations of the beats. This behavior of the gyroscope is associated with a change in its damping coefficient. Fluctuations of this coefficient with the beat frequency lead to a periodic change in the steepness of the phase-frequency characteristic of the MMG and, accordingly, to the oscillation of the measuring axis of the device regarding its original direction, approximately perpendicular to the direction of the true meridian. Metrological accuracy of the measurer reached by using the amplitude and phase gyrocompassing methods.**

Keywords—gyrocompass*, *micromechanical gyro*, *parametric excitation

I. INTRODUCTION

Micromechanical gyroscopes (MMG) are increasingly using in various fields of technology due to their small dimensions and lightweight. An overview of MMG models as devices for orientation, stabilization and navigation is given in monographs [1-2]. It was noting that, along with the undoubted advantages, the MMG has insufficiently high measurement accuracy, in connection with which the attention of researchers is focusing on the development of methods for its improvement. The analysis of publications [3-15] shows that when developing devices based on MMG, a promising approach is basing on the use of improving their characteristics by using non-traditional operating modes.

One example of such an approach is the development of a ground gyrocompass based on MMG [16]. In this work, a diagram of the device has proposed, where an MMG with a horizontal measuring axis is mounting on a rotating base. In this case, the useful signal is modulating by the angular velocity of the base, result of which it is possible to separate it from the MMG instrumental errors. Unfortunately, the author of the work does not give specific numerical estimates for improving the accuracy of the device.

Below, a new approach is proposing for determining the direction of the true meridian. The method is basing on the parametric excitation of MMG, which, as was shown by the authors earlier [17-18], can significantly increase the sensitivity of the device to the measured angular velocity, as well as expand its functionality.

An incoherent mode of parametric excitation of MMG mounted on a horizontal base is considered. A feature of this mode is the presence of "strong resonance", which increases the amplitude of oscillations of the meter rotor with respect to its resonance value, as well as "weak resonance", which reduces its amplitude of oscillations. The vibrational shape of the rotor is not preserved in this mode, and instead of harmonic vibration with one spectral component, there is a complex vibration (observed as a quasi-harmonic vibration) consisting of two spectral components with a slight difference in their frequencies. Thus, the rotor oscillations have the shape and frequency of the generated oscillations of the beats.

This behavior of the gyroscope is associated with a change in its damping coefficient. Oscillations of this coefficient with the beat frequency lead to a periodic change in the steepness of the phase-frequency characteristic of the MMG and, accordingly, to the oscillation of the measuring axis of the device relative to its original direction, approximately orthogonal to the true meridian.

II. METHOD DESCRIPTION

The specificity of the MMG operation allows based on the use of circuit solutions, without changing the design of the mechanical circuit, to increase the sensitivity of the device to the measured angular velocity due to its parametric excitation [19].

The highest sensitivity of the MMG is ensuring with conditions of the maximum amplitude of the primary oscillations, i.e. under the conditions of the implementation of the resonant tuning mode. However, the operation of MMG under conditions of a significant range of ambient temperatures, reaching 100 ° C or more, leads to a deviation of its own parameters and the frequency of the primary oscillation excitation generator, as well as to aging of the material of the sensitive element, which violates the resonant tuning.

The method under consideration is basing on modulation of the static rigidity of the MMG suspension [19]. The kinematic scheme of device is showing in Fig. 1. Modulation is providing by a slight change in the amplitude of the alternating current applied to the additional sensor winding of the torque sensor with frequency ω_m, at which a parametric excitation of MMG is creating.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

To obtain a mathematical model of a parametrically excited MMG we use the variational principle of

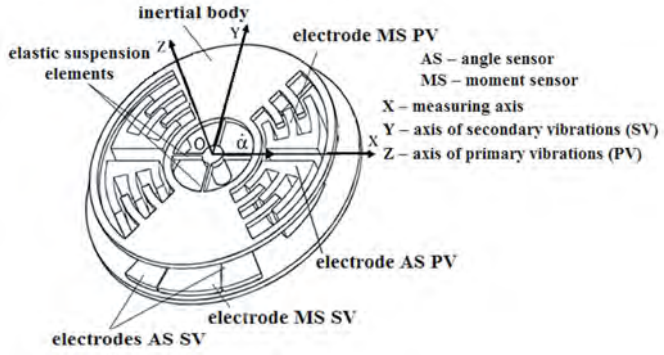


Fig. 1 The kinematic diagram of a micromechanical gyroscope

Ostrogradsky-Hamilton [20]. Differential equations of motion for the case of constant angular velocity of rotation of the base, taking into account the modulation of static rigidity relative to the axis of secondary vibrations after the factorization procedure [21] and subsequent linearization using the Jacobi matrix, will have the form:

$$\ddot{\alpha} + 2a_{\alpha}\dot{\alpha} + \omega_0^2(1 + m\sin(\omega_m t - \varphi_0))\alpha = \\ = K\omega_X \sin\Omega t, \qquad (1)$$

$a_\alpha = \frac{\mu_\alpha}{2A}$; $\omega_0 = \sqrt{\frac{k_\alpha}{A}}$; $K = \frac{(C+B-A)\,\theta_0\,\Omega}{A}$. Coefficients A, B, C are the main moments of inertia relative to the axes of the coordinate system associated with the rotor. Parameter μ_α is the coefficient of viscous friction. k_α is the stiffness of the elastic suspension relative to the axis of secondary vibrations. ω_X – the projection of the portable angular velocity of the base on the measuring axis. Ω is the excitation frequency of the primary oscillations; m is the modulation coefficient of static stiffness; θ_0 is the amplitude of the primary oscillations of the gyro rotor; φ_0 is the initial phase of the parametric excitation created by the torque sensor.

Fig. 2 shows the location of axes OX and OY relative to the true meridian. It is for the case of its resonance tuning (axis OXY) and the change in its position under incoherent mode of parametric excitation of MMG (for the case of

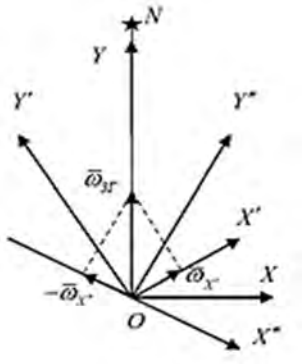


Fig. 2. Changing the position of the axes of the MMG upon parametric excitation

"strong resonance" – the position of the axes $OX'Y'$ and for the case of "weak resonance"– the position of the axes $OX''Y''$).

Note that in the case of a tuned instrument (when the position of the OX axis coincides with the direction of the West – East line), the projection of the horizontal component of the Earth's rotation $\omega_{3\Gamma}$ on the measuring axis OX is zero. With parametric excitation, a periodic change in the magnitude and sign of the $\omega_{3\Gamma}$ projection is observing. In the case of rotation of the MMG by a certain angle γ, the center of oscillation of the measuring axis OX is shifted by an amount corresponding to this angle.

It is possible to provide alignment of the axis of secondary oscillations OY with the direction of the true meridian N. This is by achieving a rotation of the device around the axis of primary oscillations OZ (so that the center of oscillation of the measuring axis OX coincides with the direction of the West – East line that is when the positive and negative values of the amplitude of oscillations of the axis OX are equal).

Along with the amplitude method for determining the direction of the true meridian, the phase method may also be used. It based on the fact, that with parametric excitation of the gyroscope, a periodic change in the damping coefficient a_α leads to a periodic change in the steepness of the phase-frequency characteristic of the MMG. In this case, the oscillation phase of the gyro rotor also periodically changes with the beat frequency relative to the resonance value equal to $-\pi / 2$. Note that for a high-quality oscillatory system, which is MMG parametrically excited, the phase-frequency characteristic in the resonance region has a significant slope. Even at small values of the modulation index m a change in the slope of the averaged phase-frequency characteristic leads to symmetrical and significant oscillations of the phase of the output signal with respect to the value of $-\pi / 2$, which makes it possible to increase the accuracy of determining the direction of the true meridian in comparison with the amplitude method.

III. THE RESULTS OBTAINED

The numerical simulation of a parametrically excited MMG (the solution of equation (1)) carried out using the Maple 9 mathematical package at the value of the beat frequency $\Omega = 0.628$ c^{-1} and at the value of modulation parameter $m = 0.00338$.

The Fig. 3 shows the time dependences of the amplitude α (solid line) and the oscillation phase φ (dotted line) of the secondary oscillations of a mounted horizontally parametrically excited MMG for various values of the angle γ of its rotation around the OZ axis, which coincides with the place vertical.

From the above solutions it follows that, along with periodic oscillations, a constant component appears that is proportional to the angle γ (blue line), both in the information signal α and in the phase value φ (t). Using this constant

component, you can automatically orientate the MMG mounted on the gyro platform in the direction of the true meridian.

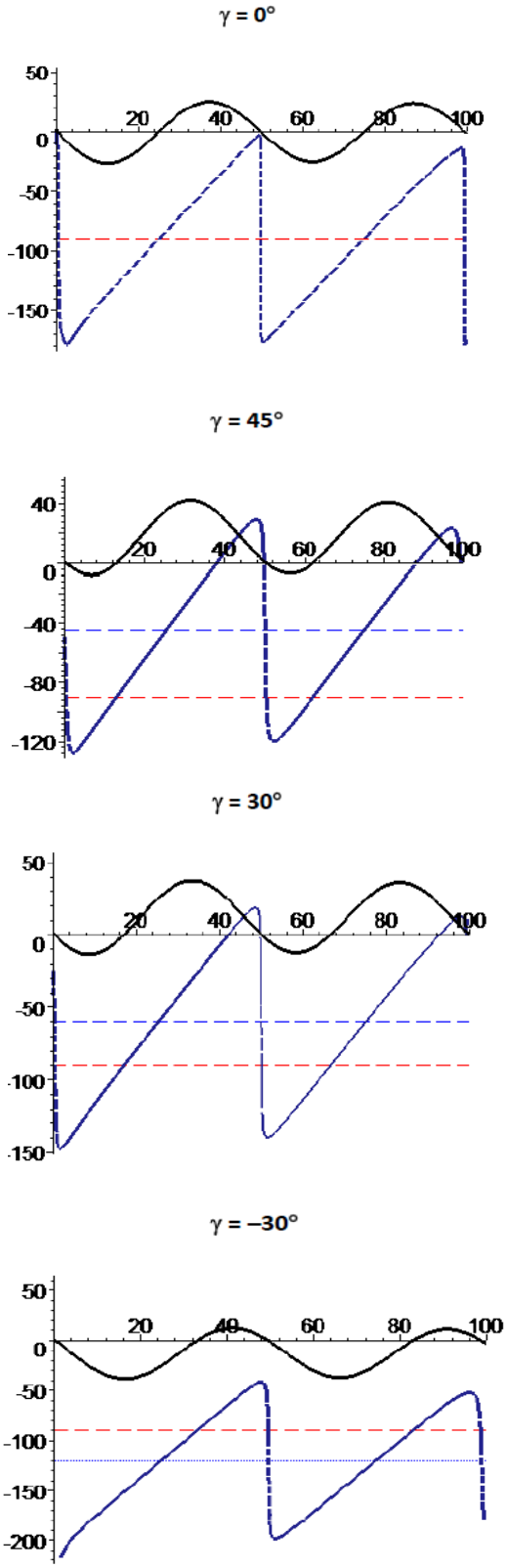


Fig. 3. Changes in the amplitude and phase of secondary oscillations of a parametrically excited MMG

IV. CONCLUSIONS

It follows from the graphs in Fig.3 that the modulation of the static stiffness of the MMG provides an increase in its sensitivity in the steady state (differential gyroscope mode) by several tens of times compared to a device in which there is no parametric excitation, which significantly increases the value of its transfer coefficient. This is due to the fact, that modulation of static stiffness significantly reduces the amount of viscous friction. This increases the duration of the linear part of the increase in the vibration amplitude, which corresponds to a significant increase in the time constant of the device and, accordingly, to a narrowing of its bandwidth.

It should be noted, that in the coherent excitation mode, an increase in the MMG sensitivity is of decisive importance for the magnitude of the phase shift between the periodically varying static stiffness of the torsion suspension and the external gyroscopic moment created by the transferred angular velocity of rotation of the device base.

An analysis of the results showed that using the parametrically excited MMG, we can determine the direction of the true meridian. Moreover, this problem can be solved both on the base of measuring amplitude fluctuations, and on the base of measuring phase oscillations.

The determination of the position of the meridian by amplitude is producing by zero values of the amplitude.

The determination of the position of the meridian by the phase of oscillations is producing by achieving equality of phase oscillations relative to the value $\varphi = -\pi/2$ (the red line in Fig. 3). It is corresponding to the position of the instrument body when the direction of the measuring axis *OX* coincides with the West - East line, and the axis of the secondary oscillations *OY* coincides with the direction of the true meridian.

The novelty of the results and conclusions is as follows:

- the new method is proposed for determining the direction of the true meridian based on the parametric excitation of a micromechanical gyroscope;
- the numerical simulation of a parametrically excited MMG was carried out;
- it was shown that the presented method allows you to automatically orientate the MMG installed on the gyro platform in the direction of the true meridian, while significantly increasing the sensitivity and quality of the device.
- the operation of the device in the parametric excitation mode allows the meter to be held in resonance mode over a wide range of operating temperatures;
- the parametric excitation mode expands the functionality of the MMG, turning it from a single-component to a two-component measurer;
- the proposed method significantly (by 1-2 orders of magnitude) increases the accuracy of measurements compared to the typical mode of operation of the device, which allows the use of MMG as an inertial measurer (for example [16]).

REFERENCES

[1] Lukyanov D.P., Raspopov V.Ya., Filatov Yu.V. Applied Theory of Gyroscopes - St. Petersburg, Electrical Appliance, 2015 – 340 p.

[2] Matveev V.V., Raspopov V.Ya. Instruments and systems of orientation, stabilization and navigation on MEMS sensors. Tula: Publishing house of TulSU, 2017 – 225 p.

[3] Vodicheva L.V., Alievskaya E.L., Koksharov E.A. and Parysheva, Yu.V. Improving the accuracy of angular rate determination for spinning vehicles – Gyroscopy and Navigation, 2012, vol. 3, no. 3, pp. 159–168.

[4] Bershtam Ya.N., Evstifeev M.I. and Eliseev D.P. Studying the alloys with high internal damping in the structure of MEMS gyro – Proc. of the 29th Conference in Memory of N.N. Ostryakov, 2014 – pp. 65–72

[5] Claire T., Guillaume G. and Mike P. High-End Gyroscopes, Accelerometers and IMUs for Defense, Aerospace & Industrial [Online] – Yole Development, 2015.

[6] Moiseev N.V. Compensation-type micromechanical gyroscope with an extended measurement range – Abstract Diss. Cand. Eng. Sc. – SPb, 2015.

[7] Eliseev D.P. Increasing Vibration Resistance of RR-type Micromechanical Gyro – Cand. Eng. Sci. Dissertation, St. Petersburg, 2015.

[8] Lestev, A.M., Combination resonances in MEMS gyro dynamics – Gyroscopy and Navigation, 2015, vol. 6, no. 1, pp. 41–44.

[9] Evstifeev, M.I., Eliseev, D.P. Improving the design of moving electrode in MEMS RR-type gyro. Gyroscopy Navig. 8, 2017 – pp. 279-286.

[10] Filatov Y.V., Boronakhin, A.M., Dao, V.B. Studying the static errors of MEMS accelerometer triad in quasiharmonic oscillation mode. Gyroscopy Navig. 8**,** 2017 – pp. 121–128.

[11] Nekrasov, Y.A., Moiseev, N.V., Belyaev, Y.V. et al. Influence of translational vibrations, shocks and acoustic noise on MEMS gyro performance. Gyroscopy Navig. 8, 2017 – pp. 31–37.

[12] Wagner, J.F. About Motion Measurement in Sports Based on Gyroscopes and Accelerometers: an Engineering Point of View – Gyroscopy and Navigation, 9, 2018.

[13] Liang, Q., Litvinenko, Y.A. & Stepanov, O.A. Method of Processing the Measurements from Two Units of Micromechanical Gyroscopes for Solving the Orientation Problem – Gyroscopy and Navigation, 9, 2018 – pp. 233–242.

[14] Bakhtieva L., Bogolyubov V. Numerical study of the three-degreed parametrically excited gyroscopic system – IOP Conference Series: Materials Science and Engineering, "11th International Conference on "Mesh Methods for Boundary – Value Problems and Applications", 2016 – p. 012014.

[15] Bakhtieva L., Bogolyubov V. Modulation of Damping in the Rotor VibratoryGyroscopes – Russ. Aeronaut., 2018 – 61: 599.

[16] Fedotov E.G. A gyrocompass based on a micromechanical gyroscope - Abstracts at the VIII Congress of Young Scientists - St. Petersburg, ITMO University, 2019.

[17] Bakhtieva L., Bogolyubov V. Parametrically excited mcroelectromechanical system in the problems of orentation of moving objects – Journal of Physics: Conference Series, 2019 – Vol. 1159, Issue 2.

[18] Bogolyubov V., Bakhtieva L. Parametrically excited microelectromechanical system in navigation problems – Proceedings of IEEE East-West De-sign & Test Symposium (EWDTS'2018), Kazan, 2018 – pp. 897-900.

[19] Golovan A.A., Makhorov G.N. and Belugin V.B. Gyroscope with parametric amplification of a useful signal on subharmonics of auxiliary movements of a sensitive element – Questions of applied mechanics, №1 (№2), Moscow, 1968 – pp. 34-38.

[20] Bakhtieva L.U., Tazyukov F.Kh. “On the stability of shells under impulse”, Uchenye zapiski Kazanskogo Universiteta. Series of Physics and Mathematics, v. 156, No. 1, Kazan, 2014, pp. 5-11.

[21] Solnitsev R.I. Computing machines in ship gyroscopes – Shipbuilding, L. – 1977.

Using Generative Adversarial Networks for Relevance Evaluation of Search Engine Results

Dmitry N. Galanin
Institute of Computational Mathematics and IT
Kazan Federal University
Kazan, Russia
trolleybus.1329@gmail.com

Nail R. Bukharaev
Institute of Computational Mathematics and IT
Kazan Federal University
Kazan, Russia
boukharay@gmail.com

Alexander M. Gusenkov
Institute of Computational Mathematics and IT
Kazan Federal University
Kazan, Russia
gusenkov.a.m@gmail.com

Alina R. Sittikova
Institute of Computational Mathematics and IT
Kazan Federal University
Kazan, Russia
sitti.alina@mail.ru

***Abstract*—In the article a new approach to the problem of relevance evaluation of the search engine results, based on generative adversarial networks (GAN), is proposed. To improve the quality of search, the generative adversarial networks are used to distinguish between relevant and irrelevant search results.**

We used a simplistic model based on fully automated reference results selection and multi-layered generator and discriminator networks with dense layers. The queries needed to generate the reference results were themselves generated by a GPT-2 like network using the same text corpus as a source, to make them potentially relevant to the search space.

The results clearly demonstrate the principal possibility and feasibility of using the described approach, despite the fact of used models being simplistic.

***Index Terms*—generative adversarial networks, machine learning, information retrieval**

I. Introduction

The problems of information retrieval are currently of considerable interest, both commercial and academic. Constant growth of processed and stored information volume causes acute problems of its structuring and cataloguing. In the Internet the extraction and selection of the necessary data is almost exclusively by means of information retrieval techniques.

Here the relevance of search engine results is the main characteristics of its quality hence competitiveness. As a measure of the importance of search engine results for its user — a human, the relevance is at some extent a subjective characteristic. However, it is obvious that it is subject to evaluation with automated methods and algorithms.

Relatively recently (mid-2010s) in the theory of neural networks, Ian Goodfellow introduced the concept of generative adversarial networks (GAN) that make possible to create objects that are more or less similar to the specified ones. This concept has found its application for a wide range of tasks — from image and photo generation to some applications in game theory. However, in the field of information retrieval, the approach described above has not yet been widely adopted.

During the last 3 years, about two dozens of papers one way or the other related to either application of neural networks in the field of information retrieval ([1]–[14]) or the relevance evaluation ([15]–[20]) were published. Unfortunately, most of the papers do not mention GANs at all, using either classical algorithms or other neural network architecture. The only paper to truly mention GANs is [13], but its author proposes the use of GAN only on the generation step, not for actual relevance evaluation.

To sum up, we assume that the approach proposed in this paper has some amount of academic novelty and provides another solution to the problem of evaluating the relevance of search engine results.

II. Terms and Definitions

A. Neural network theory

We introduce the concept of a generative adversarial network, which was coined in 2014 by Ian Goodfellow in his [21] paper.

Definition 1: Generative adversarial network (GAN) is a neural network whose main purpose is to generate objects, which are similar to the specified samples.

This behavior is implemented using the following architecture:

- the generating network G (the generator) creates (generates) objects of the specified structure.
- the discriminating network D (the discriminator) matches the generated objects with a set of reference (ground truth) values, drawing conclusions about their similarity. The G network is trained based on the feedback received from the D network (using the usual back-propagation techniques).

The GAN networks are classified as unsupervised learning nets. The pre-labeling of the training set is not required, the only essential part is the dataset of ground-truth values.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

Note that the GAN is a *concept* rather than an *architecture*, which means that the generative adversarial approach can be used with any network architecture (like multilayer perceptrons [22], LSTMs [23], etc.). In the implementation, the GAN concept is used with feedforward multilayered networks. The types of networks presented here are perceptron-like fully connected net and a net which has some drop-out layers.

B. Information retrieval theory

The parameters used below rely heavily on *term frequency* and *inverse document frequency*.

Definition 2: Term frequency indicates the significance of a particular term within the overall document.

There are multiple ways to define the term frequency [24], but the most straightforward ones are the raw count

$$\mathrm{TF}(w,d) = N_w(d) \tag{1}$$

and the normalised count

$$\mathrm{TF}(w,d) = \frac{N_w(d)}{N(d)}, \tag{2}$$

where $\mathrm{TF}(w,d)$ is the term frequency of the word w in the document d, $N_w(d)$ is the number of w's in d, while $N(d)$ is the total number of words in d (in other words, it is the length of document d in words). In the implementation provided below, both are used. The TF values provided by (1) are used in the BM25 score calculation, while (2) is used for training the GAN and subsequent predictions.

Definition 3: The inverse document frequency (IDF) is a measure of how much information the word provides, i.e., if it's common or rare across all documents.

As with TFs, there are multiple ways to define IDF. The most popular ones are based on logarithmically scaled inverse fraction of the documents that contain the word. In the definition of BM25, the IDF is most commonly defined as

$$\mathrm{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \tag{3}$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i. The main drawback of (3) is that it is negative for common words (which occur in more than half of all documents). To fix the issue, we use an offset:

$$\mathrm{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} - \log \frac{0.5}{N + 0.5}, \tag{4}$$

so the IDF of a word which occurs in *all* documents (so $n(q_i) = N$) is exactly zero, while the IDF of a word which does not occur in all the documents is strictly positive. In the implementation, the IDFs are calculated using (4).

III. Architecture

In this paper, we propose the following system architecture:

- a search query generation module designed for training the evaluating part. This module is implemented on a neural network basis, using GPT-2 like architecture (via `textgenrnn` Python package);
- a search engine module;
- a module for formalizing results that matches each query-document pair $\langle q, d\rangle$ with a set of numeric parameters $\{p_i\}$, representing characteristics of the search query result. The calculation of these characteristics is based on classical algorithms without using neural networks;
- a relevance evaluation module that accepts a set of parameters $\{p_i\}$ as input and outputs the evaluation result. This part is of the greatest scientific and practical interest. It is built with generative adversarial neural network architecture.

We will describe each module in detail separately.

A. Module for generating search queries

To train the relevance estimation module, one needs to generate a large amount of data (parameter vectors $\{p_i\}$). However, it is impractical to use random numbers for it for the reason that the probability distribution of such data may be significantly different from that obtained in real search queries. The task of researching the probability distributions of the parameters requires significant analysis, and manual generation of search queries in sufficient quantities to solve the main problem is virtually impossible. For these reasons, it is easier to generate the queries themselves using neural networks.

This sub-task can also be solved using generative adversarial networks. However, it is more promising to use the GPT-2 network architecture which was proposed in OpenAI [25]. This model was designed by the authors specifically to solve the problem of generating the texts, which are search queries in the context of this paper.

B. Search module

This part of the software system searches for a given query in the prepared database. In the simplest case (discussed here) it is just an inverted index. Nevertheless, it is possible to connect, for example, semantic tools (ontologies, thesauri, etc.). The module uses classical algorithms in the provided implementation. However, it is possible to connect neural networks to optimise search results (by storing query history determine the potentially most relevant results by analysing the similarity of queries).

C. Module for formalizing results

It is used for calculating the characteristics of the search result (the query-document pair) necessary for determining its relevance by the neural network. Among these characteristics, in particular, are parameters like term frequency in the document (of words from a query) or a combination thereof (bigrams, trigrams etc.). From a grammatical point of view it's accounting for different word forms, from the view of semantics it's necessary to account word synonyms in query context, etc. The implementation of the described module also uses only classical algorithms, without connecting neural networks.

D. Relevance evaluation module

This part of the software system is of the greatest academic and practical interest, as mentioned above. In the article, it is implemented using a generative adversarial network, which contains:

- the G subnet generates parameter vectors from the search queries. The parameters are calculated by the formalisation module. Subnet training is performed using feedback from the D subnet;
- in turn, the D subnet selects relevant results by comparing them with a dataset of known relevant queries. The source of such a dataset can be both fully manual selection of results and a mixed approach. In the latter, a certain percentage of standards is generated using classical relevance estimating functions (Okapi BM25 [26], etc.). This method allows expanding the dataset without increasing the labor cost of manual processing. However, the question concerning the quality of the results obtained requires further research. (In the implementation provided below, *all* queries were evaluated using a modified version of BM25, and the model nevertheless achieved acceptable results).

After training, the model can be used for arbitrary queries. The results obtained can be subjectively evaluated afterwards.

IV. Implementation

A. Texts used for testing

The Gutenberg Dataset [27] is used here as a corpus of texts. This is a collection of 3036 books written (in English) by 142 authors mainly of the 19th century. The dataset consists of plain-text files cleaned of any metadata, license information, and transcribers' notes, as much as possible. This condition facilitates the analysis of the texts, since it allows to get rid of the pre-processing stages.

B. The database

The database was implemented using ClickHouse database engine developed by Yandex [28]. This database engine provides fast `SELECT` queries over the tables with large (magnitude about 10^9) number of rows, which is essential to the solution to posed problem. On the other hand, ClickHouse is not suited for either `UPDATE/DELETE` queries or "heavy" `JOIN`s. However, this fact does not contribute heavily to the implementation since the base is weakly relational by its nature. So, utilising ClickHouse as the storage is a reasonable choice.

The texts were preliminarily parsed with Penn Treebank word tokeniser [29] to parse the words. These were processed by Porter stemmer [30] afterwards to retrieve word stems, leaving behind any grammatical forms. A numeric ID was assigned to each word stem occuring in the text corpus, and the IDs with the corresponding stems were stored in a table.

The inverted index has the following SQL definition:

```
CREATE TABLE default.inv_index
(
    `word_id` Int32,
    `document_id` Int32,
    `start_pos` Int32,
    `end_pos` Int32
)
ENGINE = MergeTree()
PARTITION BY document_id
ORDER BY document_id
SETTINGS index_granularity = 8192
```

So, the table is partitioned by document ID, to allow faster `SELECT` queries related to a single document (which is the frequent case upon performing the search).

The inverted indexes for bigrams and trigrams use basically the same structure except multiple `word_id` columns.

Another feature of ClickHouse which is prominently used in here is the ability to create *materialised views*. These are basically the SQL views with cached result stored on disk. They are especially useful, e.g., when querying the word counts per document (reading cached data gives a significant, sometimes order-of-magnitude, speedup).

C. The query generator

The generator uses the GPT-2 implementation with Keras using TensorFlow backend with cuDNN [31], available as a `textgenrnn` Python package [32]. The default model structure is shown on Fig. 1.

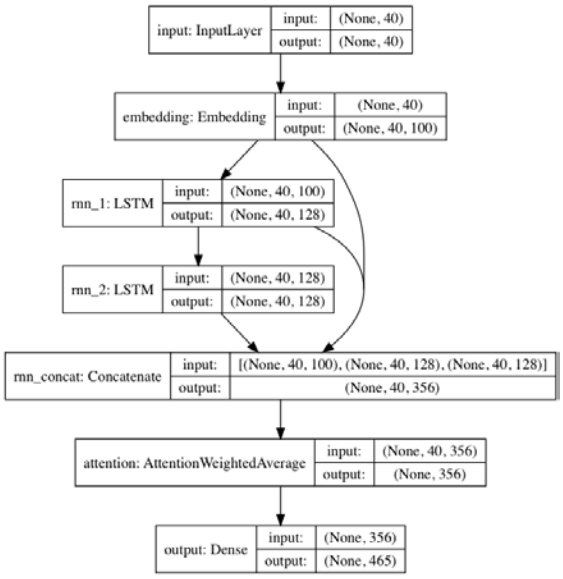


Fig. 1. The `textgenrnn` network model structure.

For the generated queries to be potentially relevant to the search space, the GPT-2 network needs to be trained on the same set of texts. To avoid longer training times while retaining the quality and relation to the domain, a dataset fraction of 5% (with a total of 152 documents and about 13 million words) was selected for training the recurrent network.

Additionally, to prevent the resulting queries to be too generic (for example, containing only the "stop words", like

articles, prepositions and otherwise commonly used words), some restrictions to the generator are imposed. Namely, a query is considered to be "good" if and only if it satisfies one of the following conditions:

1) at least 1 of the search terms (words) occurs in no more than 25% of all documents;
2) at least $\frac{1}{3}$ of the search terms (rounded to nearest integer) occur in no more than 40% of all documents;

For the task of training, a total of 242 "good" queries were selected during the generation and post-filtering steps.

Examples of generated queries include:

- *returned earl we had*
- *trouble to move in oz mode which*
- *and i don really want you to take*
- *literature seems to be the loveliest ends*
- *your distinction it is i cannot as*

The query generator was trained as a word-level model during 2 epochs. It took about 20 minutes real time per epoch when training the model on a nVIDIA® GeForce™ 2080 SUPER GPU.

D. The parameters

For the purpose of training the GAN and predicting the relevance of search results, the following parameter set was used:

1) the IDFs of each search term used in a query;
2) the term frequencies of the search terms normalised in accordance to document length (i. e. a fraction of the total number of words).
3) the term frequencies of bigrams and trigrams of the search terms.

(Note: The IDFs of bigrams and trigrams were not used to prevent the GAN from potentially devising the formula for BM25.) All vectors were padded with zeroes to a maximum length of 8 (based on an assumption that the query length is limited to 8 search terms). Hence the resulting vectors are 32-dimensional[1].

E. Selection of relevant queries

As stated above, a fully automated process for selecting the pseudo-relevant queries was used. First, the search results was ranked using the weighted BM25 function for a query $Q = q_1 q_2 \dots q_n$:

$$\begin{aligned}\text{score}(D,Q) = w_1 \sum_{i=1}^{n} \text{IDF}(q_i) \times \\ \times \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\tilde{L}}\right)} \\ + w_2 \sum_{i=1}^{n-1} \text{IDF}(q_i q_{i+1}) \times \\ \times \frac{f(q_i q_{i+1}, D) \cdot (k_1 + 1)}{f(q_i q_{i+1}, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|-1}{\tilde{L}-1}\right)} \\ + w_3 \sum_{i=1}^{n-2} \text{IDF}(q_i q_{i+1} q_{i+2}) \times \\ \times \frac{f(q_i q_{i+1} q_{i+2}, D) \cdot (k_1 + 1)}{f(q_i q_{i+1} q_{i+2}, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|-2}{\tilde{L}-2}\right)}\end{aligned} \tag{5}$$

where w_1, w_2, w_3 are the weights for single words, bigrams and trigrams respectively (in the implementation, we used $w_1 = 1$, $w_2 = 10$ and $w_3 = 100$), $f(q_1, D)$, $f(q_1 q_2, D)$ and $f(q_1 q_2 q_3, D)$ are term frequencies of a single word q_1, bigram $q_1 q_2$ and trigram $q_1 q_2 q_3$ respectively (analogously for the IDFs), $|D|$ is the document length of D, and $\tilde{L}$ is the average document length in the collection (the -1 and -2 addends for bigrams and trigrams respectively were introduced due to the fact that a document with N words trivially has $N-1$ bigrams and $N-2$ trigrams, so all the lengths are one and two less, respectively).

Second, the top 10 of the search results (ranked using (5)) was selected being treated as "relevant" (actually pseudo-relevant). In the end of the process, the total of 2158 query results[2] was selected as a ground truth dataset to train the GAN network. The results included all the TFs of single words, bigrams and trigrams, as well as the BM25 score (for reference/visualisation purpose only, the value was never used in the computations).

F. The GAN network

The network consists of the following subnets:

- the G subnet using random 100-dimensional vectors as input and 32-dimensional query result vectors as output;
- the D subnet, which accepts 32-dimensional vectors (both generated by G and the ground-truth ones) and outputs a scalar which at a glance can be interpreted as "probability" of a query being pseudo-relevant.

The activation functions mentioned hereafter are:

- the hyperbolic tangent function

$$f(x) = \tanh x; \tag{6}$$

- the sigmoid function

$$f(x) = \frac{1}{1 + e^{-x}}; \tag{7}$$

[1]Strictly speaking, it's sufficient to have vector length of 29. There are at most 7 bigrams and 6 trigrams, when the query length is limited to 8 words.

[2]This number is less than $10 \times 242 = 2420$ due to the fact that some of the queries generated less than 10 results in total.

- the leaky ReLU function [33]

$$f(x) = \begin{cases} x, & x \geqslant 0, \\ \alpha x, & x < 0. \end{cases} \tag{8}$$

The G subnet uses a sequential layer layout, as follows:

- an input layer with a dimension of 100 (for random input);
- a layer with 256 neurons, using a Leaky ReLU activation function with $\alpha = 0.2$;
- a layer with 512 neurons, using a Leaky ReLU activation function with $\alpha = 0.2$;
- a layer with 1024 neurons, using a Leaky ReLU activation function with $\alpha = 0.2$;
- an output layer with a dimension of 32, using hyperbolic tangent activation function.

The D subnet also uses a sequential layout, as follows:

- an input layer with a dimension of 32 (to feed the parameter vectors generated by G);
- a layer with 1024 neurons, using a Leaky ReLU activation function with $\alpha = 0.2$;
- a dropout layer with coefficient of 0.3;
- a layer with 512 neurons, using a Leaky ReLU activation function with $\alpha = 0.2$;
- a dropout layer with coefficient of 0.3;
- a layer with 256 neurons, using a Leaky ReLU activation function with $\alpha = 0.2$;
- a dropout layer with coefficient of 0.3;
- an output layer with a dimension of 1 (a scalar value denoting the relevance score).

The GAN network was generated using the batch size of 128, with 400 epochs. It took about 3 minutes to train the model on the same GPU mentioned above.

V. Results

To test the generated model, an additional set of 51 "good" queries were generated. Then, a classical BM25 ranking was used, after which a top-10 (pseudo-"relevant") and bottom-10 (pseudo-"irrelevant") result sets were separated. At the next stage, the results were fed through the trained model. Some of the results are shown in table I (mean values μ and standard deviations σ of GAN scores for the top-10 and bottom-10 groups).

The results demonstrate the significant difference between pseudo-relevant and pseudo-irrelevant queries, hence one can speak about the feasibility of an approach described above.

TABLE I
Query Results

Query	Relevance results	
	Top 10	***Bottom 10***
that there were no reason on lewis	$\mu = 0.8024$ $\sigma = 0.1050$	$\mu = 0.6926$ $\sigma = 0.0474$
i proposed marriage his visit	$\mu = 0.5127$ $\sigma = 0.2786$	$\mu = 0.1933$ $\sigma = 0.0694$
fanny said she	$\mu = 0.7270$ $\sigma = 0.2269$	$\mu = 0.2061$ $\sigma = 0.0086$
published it i think we found	$\mu = 0.2096$ $\sigma = 0.2077$	$\mu = 0.0822$ $\sigma = 0.0226$
i shall you heart now	$\mu = 0.4713$ $\sigma = 0.3046$	$\mu = 0.1577$ $\sigma = 0.0095$
right i had down her	$\mu = 0.4439$ $\sigma = 0.1391$	$\mu = 0.1843$ $\sigma = 0.0985$
nice declaration she got you i	$\mu = 0.7382$ $\sigma = 0.2437$	$\mu = 0.3731$ $\sigma = 0.0203$
i never moved nearer than it is	$\mu = 0.7801$ $\sigma = 0.1304$	$\mu = 0.2660$ $\sigma = 0.2241$

Nevertheless, the results showed the *principal possibility* of using GANs for the posed problem. In that case, the pseudo-relevant and pseudo-irrelevant query results were clearly separated by the neural network.

VI. Conclusions

Based on the research and the results obtained, it can be stated that the application of neural network algorithms to the problem of evaluating the relevance of information search results is promising. However, there are still issues that require further research — namely, optimization of the parameters of the applied neural networks, the organization of the knowledge base, etc.

VII. Suggestions for future research

As described above, using generative adversarial networks for evaluating the relevance of search engine results is a new technique. Hovewer, the concept was implemented in a somewhat simplified model, so it may be interesting to expand the methods with:

1) utilising more elaborate techniques which define the result of a query and their parameterisation, rather than just single words, bigrams and trigrams (for example, using word proximity metrics);
2) using semantic rather than grammatical interpretation of a query (using semantic networks, ontologies, thsauri, etc.);
3) building more sophisticated neural network structures and topologies beyond from the sequential-layered models (essentially being a multilayer perceptron-like structure).
4) using manually selected (i.e. *truly* relevant) query results as a training dataset, instead of the automatically generated ones.

References

[1] B. Mitra and N. Craswell, "Neural models for information retrieval," *CoRR*, vol. abs/1705.01509, 2017. arXiv: 1705.01509. [Online]. Available: http://arxiv.org/abs/1705.01509.

[2] S. Mohan, N. Fiorini, S. Kim, and Z. Lu, "A fast deep learning model for textual relevance in biomedical information retrieval," *CoRR*, vol. abs/1802.10078, 2018. arXiv: 1802.10078. [Online]. Available: http://arxiv.org/abs/1802.10078.

[3] Y. Chaudhary, P. Gupta, and H. Schütze, "Bionlp-ost 2019 rdoc tasks: Multi-grain neural relevance ranking using topics and attention based query-document-sentence interactions," *CoRR*, vol. abs/1910.00314, 2019. arXiv: 1910.00314. [Online]. Available: http://arxiv.org/abs/1910.00314.

[4] S. Zou, Z. Li, M. Akbari, J. Wang, and P. Zhang, "Marlrank: Multi-agent reinforced learning to rank," *CoRR*, vol. abs/1909.06859, 2019. arXiv: 1909.06859. [Online]. Available: http://arxiv.org/abs/1909.06859.

[5] D. Haddad and J. Ghosh, "Learning more from less: Towards strengthening weak supervision for ad-hoc retrieval," *CoRR*, vol. abs/1907.08657, 2019. arXiv: 1907.08657. [Online]. Available: http://arxiv.org/abs/1907.08657.

[6] R. Nogueira, "Learning representations and agents for information retrieval," *CoRR*, vol. abs/1908.06132, 2019. arXiv: 1908.06132. [Online]. Available: http://arxiv.org/abs/1908.06132.

[7] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," *CoRR*, vol. abs/1711.08611, 2017. arXiv: 1711.08611. [Online]. Available: http://arxiv.org/abs/1711.08611.

[8] L. Pang, Y. Lan, J. Guo, J. Xu, J. Xu, and X. Cheng, "Deeprank: A new deep architecture for relevance ranking in information retrieval," *CoRR*, vol. abs/1710.05649, 2017. arXiv: 1710.05649. [Online]. Available: http://arxiv.org/abs/1710.05649.

[9] R. McDonald, G.-I. Brokos, and I. Androutsopoulos, "Deep relevance ranking using enhanced document-query interactions," *CoRR*, vol. abs/1809.01682, 2018. arXiv: 1809.01682. [Online]. Available: http://arxiv.org/abs/1809.01682.

[10] C. Li, Y. Sun, B. He, L. Wang, K. Hui, A. Yates, L. Sun, and J. Xu, "NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval," *CoRR*, vol. abs/1810.12936, 2018. arXiv: 1810.12936. [Online]. Available: http://arxiv.org/abs/1810.12936.

[11] C. Zheng, Y. Sun, S. Wan, and D. Yu, "RLTM: an efficient neural IR framework for long documents," *CoRR*, vol. abs/1906.09404, 2019. arXiv: 1906.09404. [Online]. Available: http://arxiv.org/abs/1906.09404.

[12] R. K. Pasumarthi, X. Wang, C. Li, S. Bruch, M. Bendersky, M. Najork, J. Pfeifer, N. Golbandi, R. Anil, and S. Wolf, "Tf-ranking: Scalable tensorflow library for learning-to-rank," *CoRR*, vol. abs/1812.00073, 2018. arXiv: 1812.00073. [Online]. Available: http://arxiv.org/abs/1812.00073.

[13] W. Zhang, "Generative adversarial nets for information retrieval: Fundamentals and advances," *CoRR*, vol. abs/1806.03577, 2018. arXiv: 1806.03577. [Online]. Available: http://arxiv.org/abs/1806.03577.

[14] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, and X. Cheng, "A deep look into neural ranking models for information retrieval," *CoRR*, vol. abs/1903.06902, 2019. arXiv: 1903.06902. [Online]. Available: http://arxiv.org/abs/1903.06902.

[15] A. Jalilifard, V. F. Caridá, A. Mansano, and R. Cristo, "Semantic sensitive TF-IDF to determine word relevance in documents," *CoRR*, vol. abs/2001.09896, 2020. arXiv: 2001.09896. [Online]. Available: https://arxiv.org/abs/2001.09896.

[16] S. Uprety, P. Tiwari, S. Dehdashti, L. Fell, D. Song, P. Bruza, and M. Melucci, "Quantum-like structure in multidimensional relevance judgements," *CoRR*, vol. abs/2001.07075, 2020. arXiv: 2001.07075. [Online]. Available: https://arxiv.org/abs/2001.07075.

[17] C. Rosset, B. Mitra, C. Xiong, N. Craswell, X. Song, and S. Tiwary, "An axiomatic approach to regularizing neural ranking models," *CoRR*, vol. abs/1904.06808, 2019. arXiv: 1904.06808. [Online]. Available: http://arxiv.org/abs/1904.06808.

[18] D. Li and E. Kanoulas, "Active sampling for large-scale information retrieval evaluation," *CoRR*, vol. abs/1709.01709, 2017. arXiv: 1709.01709. [Online]. Available: http://arxiv.org/abs/1709.01709.

[19] S. Uprety, Y. Su, D. Song, and J. Li, "Modeling multi-dimensional user relevance in IR using vector spaces," *CoRR*, vol. abs/1805.02184, 2018. arXiv: 1805.02184. [Online]. Available: http://arxiv.org/abs/1805.02184.

[20] S. Uprety and D. Song, "Investigating order effects in multidimensional relevance judgment using query logs," *CoRR*, vol. abs/1807.05355, 2018. arXiv: 1807.05355. [Online]. Available: http://arxiv.org/abs/1807.05355.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, Montreal, Canada: MIT Press, 2014, pp. 2672–2680.

[22] F. F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65 6, pp. 386–408, 1958.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735.

[24] C. D. Manning, P. Raghavan, and H. Schütze, "Scoring, term weighting, and the vector space model," in *Introduction to Information Retrieval*. Cambridge University Press, 2008, pp. 100–123. DOI: 10.1017/CBO9780511809071.007.

[25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[26] G. Amati, "BM25," in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 257–260, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_921. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_921.

[27] S. Lahiri, "Complexity of Word Collocation Networks: A Preliminary Structural Analysis," in *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 96–105. [Online]. Available: http://www.aclweb.org/anthology/E14-3011.

[28] (). "ClickHouse - fast open-source OLAP DBMS." version 20.18, [Online]. Available: https://clickhouse.tech/ (visited on 06/23/2020).

[29] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger, "The penn treebank: Annotating predicate argument structure," in *Proceedings of the Workshop on Human Language Technology*, ser. HLT '94, Plainsboro, NJ: Association for Computational Linguistics, 1994, pp. 114–119, ISBN: 1558603573. DOI: 10.3115/1075812.1075835. [Online]. Available: https://doi.org/10.3115/1075812.1075835.

[30] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 40, pp. 211–218, 1980.

[31] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cuDNN: Efficient primitives for deep learning," *CoRR*, vol. abs/1410.0759, 2014. arXiv: 1410.0759. [Online]. Available: http://arxiv.org/abs/1410.0759.

[32] (). "Minimaxir/textgenrnn: Easily train your own text-generating neural network of any size and complexity on any text dataset with a few lines of code.," [Online]. Available: https://github.com/minimaxir/textgenrnn (visited on 06/23/2020).

[33] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *CoRR*, vol. abs/1505.00853, 2015. arXiv: 1505.00853. [Online]. Available: http://arxiv.org/abs/1505.00853.

Modeling of Smart Clothing Packet and its Porosity

Marina V. Byrdina
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
byrdinamarina@mail.ru

Mikhail F. Mitsik
Mathematics and applied informatics
Don State Technical University,
Rostov-on-Don, Russia
m_mits@mail.ru

Svetlana V. Kurenova
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
svetlana.kurenova@mail.ru

Anastasiya A. Movchun
Student, Department of
Mathematics and applied informatics
Don State Technical University,
Rostov-on-Don, Russia
anastasiya2739@mail.ru

Abstract—The paper proposes an approach to the description of the characteristics of smart clothing, which integrates with modern electronic devices and information technology. Electronic devices built into clothes help to interact with other people, with the environment, quickly find out information about your own body, measure physiological parameters. It is proposed to model a set of smart clothing as a thin flexible inextensible multilayer shell with built-in electronic devices. To design a package of "smart clothing", it is necessary to develop "smart fabrics", the creation of which requires new flexible materials for the production of electronic circuits on flexible boards, the creation of a topology of conductive threads woven into the fabric and connecting electronic devices. An approach is proposed for determining the porosity of "smart fabric" in terms of its reliability and operational safety. A model of a package of "smart clothing" based on flexible printed circuit boards, electrically conductive structures and "tencel" fabric produced using nanotechnology is proposed.

Keywords—"Smart clothing", electronic devices, "smart fabric", thin flexible inextensible shell, porosity, shell shaping

I. Introduction

Over the past ten years, a concept has emerged for light industry products, referred to as "smart clothing", which in practice means the use of electronics built into clothing and the creation of new materials [1, 2]. The appearance of such technologies is associated both with the development of microelectronics itself, and with the need to improve the properties of clothing, increase its protective properties, the reliability of its operation, comfort, convenience, etc. [3].

High level of integration of modern mobile devices together with rapid development of computer technology and related software contributes to the emergence of new areas of application of microelectronics. Just like modern mobile gadgets have replaced desktop computers in many ways, wearable clothing can have a wide range of applications and can largely replace current gadgets. Wearable electronics can be used in various areas of life: communication and communications, fashion [4, 5], entertainment, military equipment, medicine, sports, etc.

The term "Smart Clothing" means expanding the capabilities of traditional clothing. This term defines the type of woven material into which the threads are woven, providing connections between the integrated electronic devices. Microelectronic components of clothing are part of its fabric, which is outwardly invisible, immune to washing and cleaning, and does not interfere with the functionality of the clothing when it is used. In clothes not only conductive threads are used, but also embedded information input devices, antennas, sensors, as well as other components [6].

The purpose of the work is a description of approaches to the design of smart clothing, its versatility and multifunctionality, its basic characteristics, capabilities and structure, the calculation of one of the main parameters of "smart" clothing - porosity.

Research objectives:

1) developing the concept of smart clothing as a package of a multilayer flexible inextensible shell, in which each layer solves the problems of reliable operation of built-in electronic devices and ensuring the comfort of clothing;

2) calculation of porosity of smart fabric, as one of the main heat and wind protective characteristics of clothing.

The novelty of the research lies in the development of new approaches to the mathematical modeling of the package of smart clothing as a multilayer flexible inextensible shell with the properties of controlling electronic devices based on the methods of continuum mechanics and mechanics of a deformable solid body.

II. Concept of the Clothing with Integrated Electronics

From the mathematical modeling point of view, the concept of "smart clothes" can be interpreted as a multilayer flexible inextensible shell. Flexible inextensible shells are multifunctional designs that occupy a minimum of space with maximum efficiency of their use. They are used

simultaneously to set the shape, ensure strength, heat and sound insulation, and the outer layer of the structure is used as the external decoration of the interior, and thus design tasks are solved. Thin shells withstand significant loads with a minimum thickness and allow you to implement a variety of architectural forms in the design of structures of various types. Flexible shell design is a vital requirement for many types of tasks [7, 8].

One of the important tasks of modern industry is the constant concern for reducing the weight of structures while maintaining the reliability of its work. In this regard, it is necessary to consider theories of second and third order approximations, geometric and physical nonlinearity, moment theories of a deformable solid body, and also refined methods of reducing three-dimensional problems to two-dimensional ones. This includes analytical and asymptotic methods, as well as the method of successive differentiation of the relations of three-dimensional theory. Such methods should be developed not only for bodies with one small size, but also for bodies with two small sizes. When reducing three-dimensional theories to two-dimensional, it is advisable to use variational methods, which are very effective for obtaining internally consistent mathematically correct models.

It is permissible to spread methods of continuum mechanics on the mechanical behavior of matter from the macro level to the micro level. This approach is effective in explaining the behavior of materials. The field of science in which the behavior of materials with a microstructure is studied using continuous approximation methods is called generalized continuum mechanics. The development of computers allows us to solve numerically systems of equations of large dimension, which partially removes the question of the inadequacy of experiment and theory [9].

With microelectronic components integration level increase, it is possible to obtain electronic modules, data input and output data devices, indicators, power supplies, etc., which will make it possible to turn clothes into a single universal system serving a person. The concept of “smart clothing” can be interpreted as an interface between the human body and the external environment. In accordance with this, such clothing trends are developed that adapt it to external conditions. At the same time, the capabilities of “smart clothing” can be provided with new properties of the materials used.

Hundreds of different companies, universities and research centers of developed countries are currently engaged in the development of “smart clothing” concepts. Leaders of such developments are Google, Levi’s, Nike, Philips Costumer Electronics, Xiaomi, Hatsonic, Textronics Inc., Tommy Hilfiger and others. The market for smart clothing currently stands at billions of dollars.

In order to create “smart clothing”, the development of “smart fabrics” is required, which implies the creation of new flexible materials for the production of electronic circuits on flexible boards. For embedding into the fabric structure it is necessary: a flexible keyboard, flexible displays, flexible sensors. All electronic devices are interconnected by a set of conductive threads interwoven in the process of fabric formation. The topology of the conductive threads woven into the fabric is designed to ensure the connection of various electronic components in clothes made for this fabric. The material of the conductive threads is required not only to provide good electrical conductivity, but also must withstand the numerous deformations and loads that may occur during the wearing of clothes, their washing or cleaning. In this case, the material of the conductive threads should ensure reliable installation of electronic components.

The structure of “smart fabric” consists of several layers of material that form all the properties it needs. The inner layers are usually conductive, and the outer layers serve as protection from external conditions. Each fabric layer has its own functional purpose: thermal protection, waterproofing, noise absorption, vibration protection, etc.

The force of pressure on a particular point of clothing will be registered by an electronic circuit. The parameters of pressure and location are interpreted by an electronic device built into the clothing (Fig. 1).

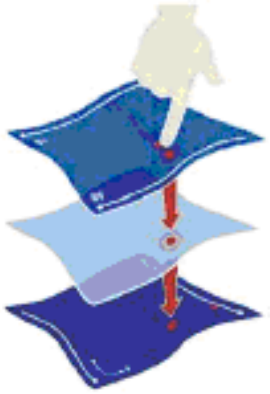


Fig. 1 – The layered structure of "smart fabric" for electronic circuit management

Thus, the new generation of materials that specialists are working on today can reconstruct the traditional idea of clothing, its functions, design and manufacturing technologies.

III. Porosity of the “Smart Closing”

An important characteristic of fabric is its porosity. Porosity is a property of a solid due to its structure and characterized by the presence of voids (pores) in it.

“Smart fabrics” just like traditional clothing fabrics are porous bodies. They contain a significant pore volume and to a lesser extent filled with fibrous material, conductive filaments and electronic devices. If we take the volume of the fabric sample as 1, then the specific porosity of the fabric can be calculated by the formula [10]

$$\omega = 1 - \frac{\gamma}{\delta}, \qquad (1)$$

where ω - specific porosity of fabric in fractions from 1; γ - average fabric density, g/cm^3 ; δ - average density of fiber, conductive filaments and electronic devices, g/cm^3 .

The porosity of “smart fabrics” varies significantly depending on their purpose. The lowest porosity are fabrics intended for the design of military equipment, for which high demands are placed on the tensile strength and wear of

the fabric, the porosity of such fabric varies from 40% to 55%. Fabrics designed to create “smart clothes” in medicine and healthcare have a sufficiently low porosity, which is due to the need for increased protection of clothing from possible infections, the presence of a large number of electronic devices and conductive threads; fabric porosity ranges from 50% to 80%. The porosity of household fabrics is quite high and varies from 75% to 90%.

Porosity is one of the basic structural characteristics of “smart fabrics”, similar to the way it is for traditional fabrics. The porosity of the fabric affects many indicators: average fabric density, strength, wear resistance, permeability, thermal conductivity, bonding efficiency and other indicators. The total porosity of the fabric consists of pores of various types: through pores; pores contained in the fibers and conductive threads; pores between the fibers and electronic devices; recesses on the front and wrong surfaces of the fabric.

The paper proposes a method for calculating fabric porosity taking into account the structure of fibers, conductive filaments and electronic devices. This method generalizes the well-known approach described in [10]. The average fabric density is calculated by the formula

$$\gamma = \frac{M_S}{1000 \cdot h} \qquad (2)$$

where: M_S - weight of 1 m^2 fabric, g; h - fabric thickness, mm.

Calculating the specific porosity of the “smart fabric” while taking into account all types of porosity can be represented as:

$$\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \frac{\gamma_E}{\delta} + \frac{\gamma}{\delta} = 1, \qquad (3)$$

Here ω_1 - specific volume of through pores; ω_2 - specific pore volume in fibers; ω_3 - specific pore volume in conductive threads; ω_4 - specific pore volume between electronic devices, fibers and conductive filaments; ω_5 - the specific volume of the recesses in the fabric on its front surface and inside out; γ_E - density of electronic devices; γ/δ - specific volume of fibrous matter.

Let V_B – specific volume of fibers in a fabric sample of fibers of 1×1 cm size. As known in the sample yarn length cumulative bases L_O and weft L_Y in centimeters, yarn diameters D_O and D_Y in centimeters, fabric thickness h in centimeters, then it’s possible to calculate the volume V_B, occupied by sample, cm^2

$$V_B = \frac{\pi D_O^2 L_O}{4h} + \frac{\pi D_Y^2 L_Y}{4h} = \frac{0{,}785}{h}\left(D_O^2 L_O + D_Y^2 L_Y\right). \qquad (4)$$

Since the specific volume of fibers V_B is calculated for sample 1×1 cm, then the specific pore volume in the fibers can be found by the formula

$$\omega_2 = V_B - \frac{\gamma}{\delta}. \qquad (5)$$

The specific volume of through pores can be calculated based on the fraction of surface fabric filling E_S

$$\omega_1 = 1 - E_S. \qquad (6)$$

The specific pore volume in the conductive threads can be calculated based on the specific volume of the threads in the sample size $1\times1\times1$ cm

$$V_H = \frac{\pi D_H^2 L_H}{4},$$

where: D_H – average diameter of the conductive thread; L_H – the total length of the conductive filament in the sample.

Then the specific pore volume in the conductive threads ω_3 is calculated by the formula

$$\omega_3 = 1 - V_H. \qquad (7)$$

The specific pore volume between electronic devices, fibers and conductive filaments can be calculated by knowing the specific volume of electronic devices in a sample of size $1\times1\times1$ cm

$$\omega_4 = 1 - V_{ED}, \qquad (8)$$

here V_{ED} – specific volume of electronic devices in the sample.

Using formulas (5 - 8) from (3) we find the specific volume of the recesses in the fabric on its front surface and the inside

$$\omega_5 = 1 - \left(\omega_1 + \omega_2 + \omega_3 + \omega_4 + \frac{\gamma_E}{\delta} + \frac{\gamma}{\delta}\right). \qquad (9)$$

It should be noted that the less through pores the fabric has, the more windproof qualities it has.

IV. Modeling of Smart Clothing Package

The design of a multilayer smart clothing package is intended to address multifunctional tasks where each of the garment layers has the desired function. An economically advantageous technology for producing a package is the production of multilayer materials by successive layering of textile fabrics with different properties and their connection into a single whole in a suitable way. This approach makes it possible to set the properties of the created layers of clothing within a very wide range, to design the surface, volumetric, hygienic and thermophysical properties of clothing, to regulate its anisotropic properties, etc. The multifunctionality of a package of smart clothing is set by the properties of each layer, taking into account the relative position.

The resulting air spaces between the layers of clothing are additional effective functional elements that allow to regulate the processes of heat transfer, as well as mass transfer through the package. Separate layers can be connected with active fillers, which increases the functionality of the package. Also, miniature devices and conductive threads are placed in the interlayers in order to regulate heat and mass transfer, to determine the medical indicators of the body, to provide protection against ionizing

radiation, from the penetration of toxic gases, vapors, microorganisms, etc. A multilayer clothing package containing electrically conductive structures is used as a protective and camouflage "smart" material; in heated clothing for special jobs; can also be used as lining for the application of various nanostructured functional materials (Fig. 2).

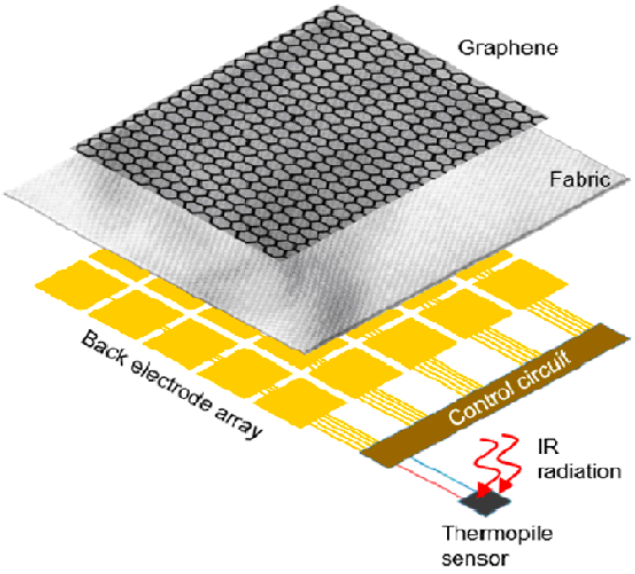


Fig. 2 – Layers of smart clothing with nanostructured materials for special tasks

A significant advantage of multilayer and multifunctional bags is determined by the fact that they have a relatively small thickness (2-3 mm) and are flexible, which makes it possible to maintain high human functionality. At the same time, the package has sufficient durability and strength, which determines the demand for smart clothes.

Electrically conductive structures in clothing are conductive threads or conductive paint. The conductive filaments can be made up of different types of filaments, such as silver, stainless steel, or carbon filaments, each with its own advantages and disadvantages. More promising for the creation of smart clothing are carbon threads, as they have high strength and heat resistance, are resistant to aggressive chemical environments and have a low density relative to metals.

Conductive paint is made on the basis of graphite powder coated onto a textile substrate by coating with a 3D printer, which is more cost effective. However, due to the crumbling of the contact tracks from the conductive paint when washing clothes, it is proposed to additionally lay conductive threads to increase the service life of flexible printed circuit boards and increase the reliability of signal reception and transmission.

Tencel is a fabric of natural origin, which is made from Australian eucalyptus wood, which is nano-processed during the production process. Tencel combines the best qualities of natural and artificial textile materials by its properties. Tencel resembles silk in texture, is as strong as viscose, and at the same time it is soft like cotton. Tencel is made from eucalyptus, so a number of properties of this tree are also attributed to the fabric, for example, bactericidal.

Tencel has a higher porosity than traditional textile materials, for example, it absorbs moisture by 50% better than cotton. In addition, due to its ability to "breathe", the tencel not only absorbs moisture, but also evaporates it, which prevents the creation of a favorable environment for the reproduction of bacteria. A person wearing clothes made of this fabric feels cool and fresh for a long time. The fabric is soft, a little shiny, like silk, but at the same time it is not slippery, warm like wool and at the same time lighter than cotton. Tencel fabric drapes well, creating soft folds, and does not crumble when cutting. Clothes made of this material are characterized by high wear resistance, respectively, they retain their spectacular appearance for a long time. Tencel fabric is characterized by high hygroscopicity and absence of static electricity, which is a very important factor for smart clothing with a conductive network.

The main advantage of using smart clothing is the constant monitoring of human body data, since the flexible printed circuit boards built into the package can provide constant non-invasive contact with the object under study, without interfering with the activities of the person. The main components of smart clothing are:

- sensors that read body parameters and generate low-intensity electrical signals;

- devices necessary for signal processing and wireless transmission;

- power supply sources;

- radio circuits providing the transmission of sensor signals to clothing components and a power source.

Conductive elements of flexible printed circuit boards must withstand the stresses to which the fabric is subjected to bending, shear and stretching during operation, as well as withstand washing the product. The advantage of forming conductivity at the tissue level is that it can be easily integrated into everyday clothing. The formation of conductivity in tissue is implemented using inorganic and organic components in microstructured and nanostructured forms.

V. Conclusions

“Smart clothing” can improve the quality of life of a person as a whole, increase the chances of survival in aggressive environments, for people with diseases and physiological abnormalities. “Smart clothing” can allow you to control the basic indicators of the body: pulse rate, respiration, body temperature, body position, etc. “Smart clothing” can help control the well-being of people working with hazardous substances in a polluted or aggressive environment. Thanks to “Smart clothing”, it is possible to remotely analyze the parameters of various technical objects, as well as provide remote medical consultation, diagnosis and treatment. Thanks to the "smart clothing" you can track the location and physical condition of the soldiers during combat missions, control the level of human fatigue. “Smart clothing” will help to create spacesuits, exoskeletons, etc., which increase the effectiveness of human actions in difficult situations.

To create “smart clothing”, it is necessary to develop a structure of “smart fabric”, which consists of several layers

of material and is modeled as a flexible inextensible multilayer shell. Important characteristics of high-tech fabric are porosity, surface density of the fabric, strength, wear resistance, permeability, thermal conductivity, bonding efficiency, etc., which can improve the protective characteristics of the fabric and its reliability in operation.

The calculation of the porosity of smart clothing in relation to a multilayer package is proposed, which will allow describing the thermophysical properties of clothing and the comfort of its operation. A description of the electrically conductive structures in a package of multilayer clothing, either in the form of conductive threads woven into textile materials, or in the form of contact tracks applied to layers of textile materials, is carried out. It has been established that the main factor in the reliable and trouble-free operation of flexible printed circuit boards manufactured on a textile basis is their stable conductivity, which ensures the reliable operation of sensors and sensors. The process of smart clothing washing might lead to the contact tracks from the conductive paint applied by the coating method to fall off, it is proposed to make additional connections using sewing lines from conductive threads to increase the reliability of the flexible printed circuit boards and increase the stability of signal transmission.

Tencel fabric is proposed for usage as a material for textile fabrics, which has a higher porosity and strength in comparison with other natural materials, which allows clothes to evaporate moisture faster, it is characterized by high hygroscopicity and the absence of static electricity, which is a very important factor for smart clothing with conductive network.

REFERENCES

[1] A.Van Halteren, R. Bults, K. Wac, N. Dokovsky, G. Koprinkov, I. Widya, D. Konstantas, V. Jones, “Wireless body area networks for healthcare: the mobihealth project,” Wearable eHealth Systems for Personalised Health Management, Studies in Health Technology and Informatics, No 108, IOS Press 2004, A. Lymberis and D. De Rossi (Eds.), pp 181 – 193.

[2] S. Park, S. Jayaraman, The wearable motherboard: the new class of adaptive and responsive textile structures, International Interactive Textiles for the Warrior Conference, 9-11 July 2002.

[3] L. Van Langenhove et al, Intelligent Textiles for children in a hospital environment, World Textile Conference Proceedings, pp. 44-48, 1-3 July 2002.

[4] "Electronic Textiles: Fiber-Embedded Electrolyte-Gated Field-Effect Transistors for e-Textiles", Wiley Online Library. John Wiley & Sons, Inc. 22 January 2009.

[5] A. Dittmar, F. Axisa, G. Delhomme, ”Smart clothes for the monitoring in real time and conditions of physiological, emotional and sensorial reactions of human,” in Proc. 25-th Annual International Conference IEEE-EMBS’03, Vol. 4, 2003, pp. 3744 – 3747.

[6] M.F Mitsik, M.V. Byrdina, L.A. Bekmurzaev. Modelimg of developable surfaces of three-dimentional geometric objects. Proceedings of 2017 IEEE East-West Design and Test Symposium, EWDTS 2017 2017. C. 8110086.

[7] L.A. Bekmurzaev, M.F Mitsik, M.V. Byrdina, G.B. Grigoryeva. Conditions of Stability of Vertical Cylindrical Soft Shell. Conference: 2018 IEEE East-West Design & Test Symposium (EWDTS). DOI: 10.1109/EWDTS.2018.8524774

[8] Extended Cold Weather Clothing System: From Wikipedia, the free encyclopedia – 2011. – URL: http://en.wikipedia.org/wiki/Extended_Cold_Weather_Clothing_System

[9] Electronics embedded in clothing-technologies and prospects. A.V. Samarin. Components and technologies. 2007. No. 4 (69). Pp. 221-228.

[10] Materials for clothing. Fabrics: textbook / B. A. Buzov, G. P. Rumyantseva. M.: ID "FORUM": INFRA-M, 2012. - 224 p. (Higher education).

Minimax Modifications of Linear Discriminant Analysis for Classification with Rare Classes

1st Kseniya Bratanova
Department of Mathematical Statistics
Kazan Federal University
Kazan, Russian Federation
bratanovakseniya@yandex.ru

2nd Iskander Kareev
Department of Mathematical Statistics
Kazan Federal University
Kazan, Russian Federation
kareevia@gmail.com

3rd Rustem Salimov
Department of Mathematical Statistics
Kazan Federal University
Kazan, Russian Federation
rustem.salimov@kpfu.ru

***Abstract*—We consider the problem of classification for imbalanced samples with rare classes. A common problem for machine learning methods in such setting is that a rare class would have extremely high classification error compared to more widespread classes. In general, this problem could be mitigated with re-sampling or fitting additional weights to control the classification errors in classes, though those methods are computationally expensive for large datasets and sometimes fail to attain appropriate results. It this paper we present cost-efficient modifications of Linear Discriminant Analysis allowing to mitigate the problem by minimizing maximal classification error among the classes. For example, this allows achieving more robust machinery malfunction detection algorithms where our expectations on recall would be more consistent among different malfunction types.**

***Keywords*—classification, imbalanced sample dataset, rare class, linear discriminant analysis, minimax error**

I. Introduction

A common issue in machine learning models when the classification problem is considered is low recall to rare classes — only a small part of samples from rare classes are recognized by a classificator. Due to the fact that most models are fitted with respect to error measures with even contribution of each sample in dataset regardless of their classes, a fitting algorithm indirectly encourages to reduce the classification errors for more frequent classes at the expense of heavily increasing classification errors for rare classes.

For example, consider an extreme situation with two classes A and B with prior probabilities of 0.99 (for A) and 0.01 (for B). Then, if the model is not good enough to clearly separate the classes (or the sample size is not big enough), the conventional algorithms would end up by classifying any input as class A giving the general classification error of 0.01. However, the recall for class B would be 0 — the resulting model wouldn't be able identify objects of class B at all! Let us note, that in many cases it would be possible to present a solution which would classify B much better at the expense of only minor reduction of recall for class A.

Such behavior is undesirable, in particular, when the rare classes are actually the target classes. Then we would be surprised to see the inability of even the most sophisticated machine learning methods to detect the target classes even when a very high general precision is reported as the result of the training. The problem is intensively researched and many approaches to reduce such negative effect in specific cases are proposed (for example, [1]–[6]).

The usual ways to mitigate the problem within machine learning framework are various forms of control of errors by cross-validation and bootstrap estimates, up-sampling, data augmentation techniques. The main disadvantages of such techniques are computational expensiveness (see [3] as example) and loose control of the error rates among classes (sometimes even small reduction of rare classes errors is not achieved). Let us observe some methods and results on the matter. Xie and Qiu [1] conducted analytical and computational investigation of performance of LDA in the setting with rare classes showing that LDA is vulnerable to the imbalanced samples. Shi, Wang, Qi, and Cheng [2] proposed modification of Logistic Regression classification methods which evaluates in a special way a Fisher discriminant for the features to improve the recall of the rare classes. Wankhade, Jonkhale, and Thool [3] considered powerful but computationally expensive rare-class-aware classification method based on the generation of ensembles of k-means classifiers with boosting and controlling the resulting errors on them. Zhang, Li, Kotagiri, Wu, and Tari [4] presented k Rare-class Nearest Neighbour classifier, which based on techniques of dynamic query neighborhoods and estimation of posterior probabilities on each of the neighborhoods.

There also families of general-purpose machine-learning methods for dealing with rare classes: Re-sampling and cost-sensitive learning. Common re-sampling methods include random oversampling and under-sampling, as well as intelligent re-sampling. The cost-sensitive learning strategy associates higher weights for incorrectly classifying samples from the rare classes. Both of those approaches might have no effect for some datasets and computationally expensive, since in most cases we need to introduce additional dummy samples and induce a series of retrainings of the classification algorithms to fit appropriately the correction weights. There have been comparative studies on the effectiveness of re-sampling and cost-sensitive learning for imbalanced classification [5], [6] showing ambiguous efficiency of the methods. Techniques for

The work was funded according to the development program of Scientific and Educational Mathematical Center of Volga Federal District, project N 075-02-2020-1478.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

data augmentation are also used to generate additional samples for rare classes (see [7] for example) with the same potential drawbacks as other re-sampling methods.

In this paper we propose a cost-effective simple modifications of Linear Discriminant Analysis classifier which are intended to minimize maximal classification error by classes. Our solution is based on the parametric nature of LDA — the input vector is supposed to have multivariate normal distribution. That allows to efficiently control the misclassification errors even with low sample size; however, due to such assumptions the performance of the method could drop dramatically whether the real distribution diverges from normal too far.

II. Definition of the LDA Modifications

A. Linear Discriminant Analysis

Suppose we observe p-variate random vector X as input data and Y is a random variable with values from $\{1, \dots, q\}$ — the class of the observation. The classification problem consists in estimating of Y based on the value of X. A machine learning model could be tuned to specific problem by fitting its internal parameters to the training dataset $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ consisting of n independent observations, and for which values $y_1, \dots, y_n$ of realizations of Y are known.

Linear Discriminant Analysis (LDA in the following) could be seen as a Bayesian solution to the classification problem in Bayesian setting where Y is assumed to be an unknown model parameter with coefficients of prior distribution $\pi_1, \dots, \pi_q$, and $X|y \sim \mathcal{N}(\mu_y, \Sigma)$ where μ_y are mean vectors and Σ is a shared covariance matrix. All the model parameters are supposed to be unknown and should be estimated based on the training sample dataset. As usual for Bayesian solutions, LDA decision rule could be expressed as maximizing the posterior probability:

$$\hat{y}(x) = \arg \max_{y \in \{1,\dots,q\}} \mathbf{P}(Y = y | X = x), \tag{1}$$

where the posteriors are calculated with estimates $\mu_y = \hat{\mu}_y$, $\Sigma = \hat{\Sigma}$. See [8] for more details on LDA. Let us note, that by the form, LDA is an optimal classifier if the underlying assumptions on the distributions of X and Y are true.

B. Modifications for Minimizing Maximal Error

In this paper we consider three similar modifications. Essentially, all of them consist in bringing in additional weights to (1). Then the values of the weights are fitted to minimize the maximal error:

$$\mathcal{E}(\hat{y}) = \max_{k \in \{1,\dots,q\}} e_k(\hat{y}), \tag{2}$$

where $e_k = \mathbf{P}\big(\hat{y}(X) = Y | Y = k\big)$.

Here are the proposed modifications:

- For LDA+ we change decision rule (1) to be:

$$\hat{y}^*(x; w) = \arg \max_{y \in \{1,\dots,q\}} \log \mathbf{P}(Y = y | X = x) + w_y,$$

where $w = (w_1, \dots, w_q)$. Let us note that $LDA+$ is actually still LDA with altered prior distribution coefficients.

- For LDA* the rule (1) becomes:

$$\hat{y}^*(x; w) = \arg \max_{y \in \{1,\dots,q\}} w_y \log \mathbf{P}(Y = y | X = x),$$

where $w = (w_1, \dots, w_q)$.

- For LDA+* the rule (1) becomes:

$$\hat{y}^*(x; w) = \arg \max_{y \in \{1,\dots,q\}} w_y \log \mathbf{P}(Y = y | X = x) + w^*_y,$$

where $w = (w_1, \dots, w_q, w^*_1, \dots, w^*_q)$.

For all the modifications the values of weights are fitted as minimizing (2):

$$w = \arg \min_w \mathcal{E}\big(\hat{y}^*(\cdot; w)\big) = \arg \min_w \max_k e_k\big(\hat{y}^*(\cdot; w)\big).$$

We solved this minimax problem numerically using general-purpose optimization algorithms. Let us note, that e_k are expressed through a p-variate integrals over complicated areas. Thus, direct calculation of e_k is a computationally expensive process. Instead, we estimated the integrals using Monte-Carlo simulations. Our tests showed that even 100 Monte-Carlo samples are enough to achieve sufficient precision for $p \le 10$.

As seen from the definition, the proposed modifications increase the computational complexity only by number of features p. It is an attractive property of our method compared to the common ways for improvement of rare classes recalls, such as sample and class weighting or data augmentation. For those methods the computational complexity increases both in p and the sample size n which makes it expensive to be used in larger datasets. As an example, we considered cross-validated minimax fitting of Random Forrest and Decision Tree, where the weights of classes were fitted to minimize the maximal class classification error based on the cross-validated error estimate by classes. Let us note, that not only LDA+, LDA*, and LDA+* fitted up to several times faster, but also presented better performance in terms of the maximal classification errors (see Tables IV, VI, VIII, and X) than the aforementioned cross-validated minimax fitting.

III. Results of Computational Experiments

All the computational algorithms were coded on Python programming language and run on the corresponding CPython platform [13]. For the standard machine learning methods of Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and LDA we used the implementations from the Python package scikit-learn [14].

A. Simulations with Normal Distribution

As the first part of the numerical evaluation of the performance of the presented modifications LDA+, LDA*, LDA+*, we provide the results of the simulations for X with multivariate normal distribution with $p = 5$, $q = 3$. The investigation consisted of 100 independent experiments. In each experiment the mean vectors and the covariance matrix were generated from Gamma distributions with the parameters $(2, 2)$, $(2, 3)$, $(2, 3)$ for the mean vectors and $(2, 2)$ for the covariance

matrix. For each experiment the sample of size 1000 was drawn for 1st class, 100 for 2nd, 10 for 3rd resulting in highly imbalanced samples. Each of the methods LDA, LDA+, LDA*, and LDA+* was trained on the training part of the sample dataset and their maximal error was computed on the testing parts. The obtained sample maximal errors were then averaged over the 100 independent experiments. Due to the fact that LDA is nearly optimal for normally distributed data, we didn't consider other machine learning classifiers here.

The results of simulations are presented in Table I. There $\max_k e_k$ denotes the maximal classification error. As one could expect, LDA+, LDA*, LDA+* perform much better in terms of the maximal error. Another expected result is the best performance of LDA+ due to the normality of the data. An example of fitted weights values is contained in Table II.

TABLE I
CLASSIFICATION ERROR ON SIMULATED EXPERIMENTS

Model name	$\max_k e_k$
LDA	0.426
LDA+	0.220
LDA*	0.273
LDA+*	0.299

TABLE II
EXAMPLE OF WEIGHTS VALUES FOR MODIFICATIONS OF LDA

Model name	**weight 1**	**weight 2**	**weight 3**
LDA+	0.609	3.147	1.078
LDA*	1.002	1.009	1.003
LDA+*	0.998	1.004	1.003
(w^* *weights*)	1.008	1.014	1.007

B. Evaluation on Real-Life Datasets

During this part of the investigation we considered 4 datasets. For each dataset we evaluated performance of several conventional classifiers using 5-fold cross-validation (4 folds for a training and 1 for a testing subsets). To compare with our proposed models LDA+, LDA*, LDA+*, we considered the following models:

- Logistic Regression without regularization, with L1 regularization, with L2 regularization.
- Decision Tree and Random Forest.
- K-Nearest Neighbors.
- Ordinary LDA.

See [8] for more details and implementation details on the listed methods.

In addition to that we considered Decision Tree and Random Forest with the minimax fitting by cross-validation. The minimax fitting by cross-validation consisted in bringing in weights for classes which were fitted to minimize the maximal cross-validated error estimate for the classes. The control of the test error was done by introducing validation fold, so for those two methods and LDA+ we used 5-fold cross-validation with 3 folds for the training, 1 fold for the validation, and 1 fold for the testing.

For each comparison we considered 4 types of error metrics:

- $\max_k e_k$ — maximal classification error (maximal recall).
- E^{10}, E^5, E^2, where $E^r = \sqrt[r]{\sum_{k=1}^{p}(e_k)^r/p}$ — intermediate error measures between maximal error and general error measures. We introduced these metrics to see the overall efficiency of the methods in more details, to see effect on the other classes, not only the worst-classifiable.

For each type of error in tables we marked the lowest error rate by bold font.

1) Glass Type Classification Dataset: This dataset [9] contains several attributes of a glass with type of the glass as response. For this dataset $p = 9$, $q = 6$. The number of samples among the classes: 70, 76, 17, 13, 9, 29. The results of the evaluation are presented in Table III. Table IV contains the analogous evaluation in comparison with Decision Tree and Random Forrest with minimax fitting by cross-validation as described in the beginning of this section.

For this dataset our modifications show the least minimax error compared to the other classifiers. However, the advantage compared to Random Forest is insignificant, and the error is greater for more relaxed metrics E^5 and E^2. To summarize, Random Forrest seems more preferable choice for this dataset. Let us also note, that Random Forest and Decision Tree show by far the best minimax errors compared to standard classification methods under the consideration. As seen in Table IV, Random Forest doesn't benefit from additional fitting of class weights to minimize maximum error. However, Decision Tree becomes much more sensitive to the rare classes allowing it to outperform our modifications of LDA in E^2 metric.

TABLE III
CLASSIFICATION ERRORS FOR GLASS TYPE DATASET

Model name	$\max_k e_k$	E^{10}	E^5	E^2
Logistic Regr.	1.00	0.921	0.875	0.774
Logistic Regr., L1 regul.	1.00	0.846	0.756	0.629
Logistic Regr., L2 regul.	1.00	0.852	0.775	0.654
Decision Tree	0.77	0.663	0.597	0.494
K-Nearest Neighbors	0.90	0.777	0.699	0.580
Random Forest	0.70	0.586	**0.506**	**0.409**
LDA	1.00	0.839	0.732	0.595
LDA+	**0.60**	0.559	0.534	0.489
*LDA**	**0.60**	0.558	0.535	0.489
*LDA+**	**0.60**	**0.541**	0.513	0.470

TABLE IV
CLASSIFICATION ERRORS FOR GLASS TYPE DATASET WITH MINIMAX FITTING BY CROSS-VALIDATION

Model name	$\max_k e_k$	E^{10}	E^5	E^2
Decision Tree	0.72	0.616	0.563	**0.501**
Random Forest	0.79	0.705	0.655	0.577
LDA+	0.70	0.625	0.586	0.533
*LDA**	**0.68**	**0.513**	**0.560**	0.594
*LDA+**	0.70	0.537	0.588	0.626

2) Online Shopper's Intention Dataset: In this dataset [10] we supposed to guess whether a client would buy given

product based on its attributive description. For this dataset $p = 15$, $q = 2$. The number of samples among the classes: 10422, 1908. The results of the evaluation are presented in Table V. Table VI contains the analogous evaluation in comparison with Decision Tree and Random Forrest with minimax fitting by cross-validation.

We see that our modifications of LDA significantly outperform all other methods in terms of the rare classes classification errors. Another notable fact is that Decision Tree shows better recall for the rare classes than Random Forrest.

TABLE V
CLASSIFICATION ERRORS FOR ONLINE SHOPPER'S INTENTION DATASET

Model name	$\max_k e_k$	E^{10}	E^5	E^2
Logistic Regr.	0.985	0.919	0.858	0.697
Logistic Regr., L1 regul.	0.642	0.599	0.558	0.454
Logistic Regr., L2 regul.	0.645	0.602	0.561	0.456
Decision Tree	0.476	0.444	0.415	0.343
K-Nearest Neighbors	0.712	0.664	0.620	0.504
Random Forest	0.505	0.471	0.440	0.358
LDA	0.649	0.606	0.565	0.459
LDA+	**0.364**	**0.340**	**0.317**	**0.263**
*LDA**	0.405	0.378	0.353	0.299
*LDA+**	0.393	0.366	0.342	0.289

TABLE VI
CLASSIFICATION ERRORS FOR ONLINE SHOPPER'S INTENTION WITH MINIMAX FITTING BY CROSS-VALIDATION

Model name	$\max_k e_k$	E^{10}	E^5	E^2
Decision Tree	0.477	0.445	0.415	0.344
Random Forest	0.530	0.495	0.462	0.376
LDA+	0.400	0.373	0.348	**0.287**
*LDA**	**0.383**	**0.358**	**0.335**	0.295
*LDA+**	0.408	0.381	0.356	0.305

3) Red Wine Quality Dataset: In this dataset [11] the intention is to automatically classify quality of wine based on severl attributes. For this dataset $p = 10$, $q = 6$. The number of samples among the classes: 681, 638, 199, 53, 18, 10. The results of the evaluation are presented in Table VII. Table VIII contains the analogous evaluation in comparision with Decision Tree and Random Forrest with minimax fitting by cross-validation.

Again, our modifications of LDA are better for all the metrics under consideration. The additional minimax fitting by cross-validation make the metrics worse for Decision Tree and Random Forest. That seems to have happened because of the reduced sample size due to additional fold in cross-validation.

4) Sloan Digital Sky Survey DR14 Dataset: The dataset [12] consists of observations of space taken by the SDSS. Every observation is described by several feature columns and 1 class column which identifies it to be either a star, galaxy or quasar. For this dataset $p = 10$, $q = 3$. The number of samples among the classes: 4998, 415 (artificial reduction of sample size was done to increase imbalance), 850. The results of the evaluation are presented in Table IX. Table X contains the analogous evaluation in comparison with Decision Tree and Random Forrest with minimax fitting by cross-validation.

TABLE VII
CLASSIFICATION ERRORS FOR RED WINE QUALITY DATASET

Model name	$\max_k e_k$	E^{10}	E^5	E^2
Logistic Regr.	1.00	0.960	0.923	0.848
Logistic Regr., L1 regul.	1.00	0.943	0.903	0.820
Logistic Regr., L2 regul.	1.00	0.946	0.906	0.823
Decision Tree	0.90	0.767	0.715	0.672
K-Nearest Neighbors	1.00	0.937	0.895	0.834
Random Forest	1.00	0.848	0.775	0.699
LDA	1.00	0.890	0.829	0.739
LDA+	0.85	0.723	**0.666**	**0.617**
*LDA**	0.85	0.729	0.675	0.631
*LDA+**	**0.80**	**0.716**	0.678	0.643

TABLE VIII
CLASSIFICATION ERRORS FOR RED WINE QUALITY WITH MINIMAX FITTING BY CROSS-VALIDATION

Model name	$\max_k e_k$	E^{10}	E^5	E^2
Decision Tree	0.94	0.806	0.751	0.703
Random Forest	1.00	0.852	0.786	0.715
LDA+	0.80	0.708	0.669	0.634
*LDA**	**0.75**	**0.643**	**0.606**	**0.583**
*LDA+**	**0.75**	0.657	0.626	0.600

For Sloan Digital Sky Survey dataset our modifications dramatically outperform the standard machine learning methods. Additional minimax fitting by cross-validation improve the results for Decision Tree and Random Forrest, however, not much enough.

TABLE IX
CLASSIFICATION ERRORS FOR SLOAN DIGITAL SKY SURVEY DR14 DATASET

Model name	$\max_k e_k$	E^{10}	E^5	E^2
Logistic Regr.	0.999	0.933	0.894	0.790
Logistic Regr., L1 regul.	0.713	0.639	0.573	0.415
Logistic Regr., L2 regul.	0.735	0.658	0.590	0.427
Decision Tree	0.496	0.445	0.399	0.299
K-Nearest Neighbors	0.824	0.774	0.743	0.660
Random Forest	0.598	0.535	0.479	0.352
LDA	0.431	0.386	0.346	0.253
LDA+	0.080	0.076	0.073	0.068
*LDA**	**0.069**	**0.065**	**0.063**	**0.059**
*LDA+**	0.083	0.076	0.073	0.069

TABLE X
CLASSIFICATION ERRORS FOR SLOAN DIGITAL SKY SURVEY DR14 DATASET WITH MINIMAX FITTING BY CROSS-VALIDATION

Model name	$\max_k e_k$	E^{10}	E^5	E^2
Decision Tree	0.460	0.412	0.370	0.287
Random Forest	0.472	0.423	0.379	0.293
LDA+	0.086	0.078	0.074	0.071
*LDA**	0.073	**0.067**	**0.065**	**0.063**
*LDA+**	**0.082**	0.076	0.073	0.069

IV. CONCLUSION

We considered three modifications of LDA with intend to get classifier with minimax error over the classes. The numerical experiments on the simulated and the real-life datasets

showed surprisingly high efficiency of the modifications compared to even very powerful method of Random Forest. For some datasets (Glass Type Classification, for example) we might say that our modifications are better in terms of minimizing the maximal error, however become worse in terms of even a little more relaxed error measures. For some others (Sloan Digital Sky Survey DR14 Dataset, for example) we have seen surprisingly high efficiency of the proposed methods even in comparison with Random Forest and more relaxed error measures.

As with comparison between LDA+, LDA*, LDA+* themselves — the experiments haven't unambiguously shown the best of them. They gave similar results, and it is not clear which one is the most efficient in general. So far the difference in results seems to be insignificant.

In the future investigations on the topic we might try to find analytical simplifications on the problem of optimization of the weights for minimizing the maximal classification error. That would allow to reduce the computational cost even further and, maybe, to achieve even greater accuracy of the methods.

REFERENCES

[1] J. Xie, Z. Qiu, "The effect of imbalanced data sets on LDA: A theoretical and empirical analysis," Pattern recognition, vol. 40, no. 2, pp. 557–562, February 2007.

[2] B. Shi, J. Wang, J. Qi, Y. Cheng, "A novel imbalanced data classification approach based on logistic regression and Fisher discriminant," Mathematical Problems in Engineering, January 2015.

[3] K. Wankhade , K. Jondhale , V. Thool, "A hybrid approach for classification of rare class data, " Knowledge and Information Systems, vol. 56, no. 1, pp. 197–221, July 2018.

[4] X. Zhang , Y. Li, R. Kotagiri, L. Wu, Z. Tari, M. Cheriet, "KRNN: k Rare-class Nearest Neighbour classification," Pattern Recognition, vol. 62, pp. 33–44, February 2017.

[5] C. Elkan, "The foundations of cost-sensitive learning," Proceedings of IJCAI, pp. 973–978, 2001.

[6] G.M. Weiss, K. McCarthy, B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs," Proceedings of ICDM, pp. 35–41, 2007.

[7] S. Beery , Y. Liu, D. Morris, J. Piavis, A. Kapoor, N. Joshi, M. Meister, P.. Perona, "Synthetic examples improve generalization for rare classes," The IEEE Winter Conference on Applications of Computer Vision, pp. 863–873, 2020.

[8] T. Hastie, R. Tibshirani, J. Friedman, "The elements of statistical learning: data mining, inference, and prediction," Springer Science & Business Media, 2009.

[9] Glass Classification Dataset, URL: https://www.kaggle.com/uciml/glass

[10] Online Shopper's Intention Dataset, URL: https://www.kaggle.com/roshansharma/online-shoppers-intention

[11] Red Wine Quality Dataset, URL: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

[12] Sloan Digital Sky Survey DR14 Dataset, URL: https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey

[13] Python, URL: https://www.python.org/

[14] scikit-learn, URL: https://scikit-learn.org/

Smart Fabric Thermal Conductivity Modeling

Mikhail F. Mitsik
Mathematics and applied informatics
Don State Technical University,
Rostov-on-Don, Russia
m_mits@mail.ru

Marina V. Byrdina
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
byrdinamarina@mail.ru

Svetlana V. Kurenova
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
svetlana.kurenova@mail.ru

Dmitry B. Kelekhsaev
International logistics systems and complexes
Platov South-Russian State
Polytechnic University (NPI)
Novocherkassk, Russia
d-kelekhsaev@mail.ru

***Abstract*—The paper proposes an approach to constructing a mathematical model of the smart fabric thermal conductivity, which are integrated with modern electronic devices, information technologies and new materials. The initial and boundary conditions for the heat equation are described, which simulate various conditions at the fabric boundary, as well as between its layers, and on the basis of which a unique solution is found for the heat conduction problem. It is shown that the average temperature of the fabric can be determined as an integral characteristic over the entire surface. The solution to the boundary value problem on the temperature distribution is obtained numerically on the basis of finite-difference schemes. The results of the experimental values of temperatures agree with the calculated values with an error of no more than 7%. The paper is the basis for solving the problem of thermal conductivity for smart fabric.**

Keywords—heat conduction problem, initial and boundary conditions, smart fabric, finite-difference scheme, electronic devices, thin flexible inextensible shell.

I. INTRODUCTION

Currently, clothing is subject to a high level of hygienic, technological and aesthetic requirements, which together should increase the quality of life. Smart fabrics in the near future will be used both for domestic and technical, medical and other purposes [1].

The processes of heat exchange between a person and the environment have an important role in human life and therefore the heat-protective functions of clothing are of particular importance [2]. From numerous studies by various authors it follows that thermal comfort has a beneficial effect on human health, positively affects labor processes, they are more productive and less tiring, rest is also more effective. In conditions of thermal comfort, the physiology of the body and thermoregulation work with less load, which reduces the risk of colds and other diseases [3, 4].

The clothing industry is often faced with the need to choose materials for clothing empirically, not taking into account hygiene requirements, climatic conditions, labor characteristics and many other factors. The design of smart heat-protective clothing in relation to specific climatic conditions and production requirements is a large and very complex scientific problem that can be solved by creating new smart materials and technologies, taking into account the physiological data of a person, climatic and industrial conditions [5]. It should be noted that at present microclimate parameters are determined by norms SanPin 2.2.4.548 – 96, SP 60.1330.2012.

The purpose of the paper is to develop general approaches to determine the heat-shielding properties and thermophysical characteristics of the fabric, methods for calculating the thermal conductivity of smart fabric depending on its heat-shielding characteristics and environmental conditions.

Research objectives:

1) describing the heat transfer processes that occur between the human body, layers of clothing and the external environment in relation to smart clothing;

2) building a mathematical model of the smart fabrics thermal conductivity based on the heat balance equation of the human body;

3) formulating the initial and boundary conditions for the heat equation to find the only solution to the problem that satisfies real physical processes;

4) calculating the temperature distribution over the shell thickness using a model example and comparing the calculated temperatures with experimental data.

The novelty of the research lies in the statement of the problem of thermal conductivity for smart fabrics based on the heat balance equation and a description of the initial and boundary conditions for the heat equation for various layers of the smart fabric package.

II. MATHEMATICAL MODEL OF SMART FABRICS THERMAL CONDUCTIVITY

Heat transfer between the human body, layers of clothing and the external environment is complex and depends on physiological conditions and processes, climatic conditions, as well as external factors. Heat transfer is determined by the properties of clothing materials [6, 7]. The characteristics of the heat-shielding properties of smart fabric depend on the thermal conductivity of its constituent fibers, threads and microelectronic devices, their shape, as well as on how they fill the volume of fabric. A fabric

having a greater porosity has less thermal conductivity, since a large fraction of the volume in it is filled with air, which conducts less heat than any fabric.

Factors that influence the thermal resistance of materials, fabrics and electronic devices include: thickness of fabric elements, density, moisture, fiber types, type and quality of fibrous material, fabric structure, bulk weight, breathability, etc.

The heat-shielding properties of clothing are a poorly understood area. It is necessary to create a common methodology for determining the heat-shielding properties and thermophysical characteristics of clothes, the integrated use of new materials and devices woven into clothes, which allow us to analyze the physiological state of a person and quickly display information about basic vital signs. Thermal properties of clothing can be considered as the main operational indicator.

One of the defining properties are the thermophysical properties of "smart" fabrics, and, in particular, their thermal conductivity. To assess the heat-shielding properties of fabrics, indicators of the following characteristics are used: thermal conductivity, thermal resistance, heat capacity, thermal diffusivity. It is known that fabrics mainly represent a duaxial weave network, which is a porous system with structural characteristics determined by the manufacturing method.

Heat transfer in porous materials occurs using conductive heat transfer along the weave network itself, as well as convection due to the movement of air (or liquid) through the pores of the material. These processes are interconnected, since a complex convective and radiant heat exchange occurs between the weave network and the air. Accordingly, to find the value of the heat flux moving through the porous material, the combined thermal conductivity coefficient λ. is used. Coefficient λ. is conditional and determines the thermal conductivity of some homogeneous material, such that with the same shape, size and temperature at the borders, the same amount of heat passes through it as through a given porous material.

The creation of smart fabrics with desired thermophysical properties is one of the main tasks for technologists in the future. To solve this problem, it is necessary to obtain dependencies that will allow us to determine the structural characteristics of fabrics for given values of thermophysical parameters.

Modeling of the thermal conductivity of fabrics is determined from the condition of thermal comfort of a person, for which a large number of experimental studies [8] were conducted. accordingly, traditional methods for calculating the thermal conductivity of fabric are modeled on the basis of the heat balance equation.

$$Q_S = Q_M - (Q_W + Q_E + Q_R + Q_C), \qquad (1)$$

where $Q_S -$ heat storage of the human body,

$Q_M -$ metabolic heat, $Q_W -$ thermal equivalent of mechanical work, $Q_E -$ heat emission by evaporation,

$Q_R -$ radiant heat, $Q_C -$ convective heat.

The terms of equation (1) were obtained experimentally, as well as on the basis of well-known theoretical principles [8].

To create an adequate model of thermal conductivity of smart fabric, we make the following assumptions. We divide the entire surface of the "smart" fabric into elementary sections, on each of which the fabric will have constant parameters of thermal resistance, thickness, specific heat density, thermal diffusivity. This can be achieved by dividing the entire surface into zones with a duaxial weave network, fabric zones containing conductive threads, and zones containing microelectronic devices.

Then on the surface of the elementary section of the partition, we can assume that the temperature on the surface of the fabric changes only in the direction perpendicular to the surface. Accordingly, instead of the general three-dimensional heat equation, we obtain a one-dimensional equation, which greatly simplifies the statement of the problem.

Under the above assumptions, equation (1) implies the one-dimensional unsteady heat equation

$$c\rho\frac{\partial T}{\partial t} = \frac{\partial}{\partial x}\left(\lambda\frac{\partial T}{\partial x}\right) - \frac{\alpha}{h}(T - T_O) + Q(x,t), \qquad (2)$$

where: $c -$ specific heat, $\rho -$ material density,

$T -$ temperature, $t -$ time,

$x -$ current coordinate on the surface of the fabric, reporting perpendicular to the surface, *m*,

$\lambda -$ coefficient of thermal conductivity, $\frac{J}{m \cdot s \cdot K}$,

$\alpha -$ heat transfer coefficient between the environment and the surface of the fabric, $h -$ material thickness, *m*,

$T_O -$ environment temperature, $Q(x,t) -$ heat sink intensity.

Equation (2) for the classification of partial differential equations of the second order are of parabolic type [7].

Coefficients c, ρ, λ and α will depend significantly on the area under consideration, so the heat equation should be solved separately for each area. It should be noted that the obtained temperature distribution field must be safe for the reliable operation of the gadgets built into the clothes.

III. Initial and Boundary Conditions for the Heat Equation

The heat equation itself has infinitely many solutions. The problem of solving the heat equation is that it is necessary to find the only solution that satisfies the given initial and boundary conditions, while the solution should simulate a real physical process [9]. The initial condition is the temperature distribution inside the fabric

$$T = T(x, t_0), \qquad (3)$$

where the value is usually taken as the reference time of the process under study.

The boundary conditions for the unsteady heat equation are the temperature distribution conditions at the boundaries of each elementary region. In this case, the boundary and initial conditions should simulate the designed thermophysical conditions.

Since smart clothing can be interpreted as a multilayer flexible inextensible shell, depending on the thermophysical conditions under consideration, boundary conditions of the first, second, third, or fourth kind can be set by layers of fabric.

A boundary condition of the first kind sets the temperature field on the surface of the fabric, which is a known function of time

$$T(M,t) = f(M,t), \qquad (4)$$

here $M = M(x, y, z)$ – point on the surface of the fabric.

At each elementary site, function (4) takes a constant value, however, it changes when moving to another site.

A boundary condition of the second kind sets the value of the heat flux at each point of the elementary section

$$\lambda \frac{\partial T(M,t)}{\partial n} = q(M,t), \qquad (5)$$

where $q(M,t)$ – heat flux at a point M, partial derivative of temperature at a point M calculated in the direction of the normal vector to the fabric surface.

Using a boundary condition of the second kind, one can simulate the effect on the temperature of a fabric of radiant energy emanating from an external source on the surface of an opaque fabric.

A boundary condition of the third kind sets the relationship between the temperature of the fabric on the surface and the magnitude of the heat flux through the fabric

$$\lambda \frac{\partial T(M,t)}{\partial n} = \alpha(T(M,t) - T_O), \qquad (6)$$

where is the partial derivative of the temperature at the point M is calculated in the direction of the normal vector to the fabric surface.

A boundary condition of the third kind describes the transfer of heat by convection from fabric to the environment - air having a temperature T_O.

Based on a boundary condition of the third kind, it is possible to simulate the radiant heat transfer between two layers of fabric with temperatures on the surface T_1 and T_2, and accordingly, from T_2 to T_1

$$Q_{rez} = \sigma(T_2^4 - T_1^4), \qquad (7)$$

here: Q_{rez} – the value of the resulting heat flux during heat transfer,

σ – Stefan - Boltzmann constant.

Then, the amount of flow during heat transfer to the first fabric is

$$\lambda \frac{\partial T_1(M,t)}{\partial n} = \sigma(T_2^4 - T_1^4). \qquad (8)$$

In this case, the heat transfer coefficient α is a function depending on the temperature of the fabric.

A boundary condition of the fourth kind defines the relationship between the heat flux and the temperature at the boundary of two fabric layers in perfect contact

$$\lambda_1 \frac{\partial T_1(M,t)}{\partial n} = \lambda_2 \frac{\partial T_2(M,t)}{\partial n}. \qquad (9)$$

where λ_1, λ_2 – heat conductivity coefficients of the first and second layer in a given elementary section.

A boundary condition of the fourth kind models the amount of heat that is taken from the first layer to its boundary, the same amount of heat is transferred to the second layer with perfect thermal contact [10].

If, however, a phase transition occurs at the boundary of two layers, then the boundary conditions will change taking into account the absorption, or the release of the heat of the phase transition

$$\lambda_1 \frac{\partial T_1(M,t)}{\partial n} - \lambda_2 \frac{\partial T_2(M,t)}{\partial n} = \rho Q_1 U, \qquad (10)$$

where $U = \frac{\partial x_n}{\partial t}$ – velocity of the phase transition at the interface between the layers, Q_1 – phase transition heat.

It should be noted that the boundary and initial conditions of the problem, in turn, are approximate and, accordingly, are approximated when the heat conduction problem is formulated. Errors inherent in the formulation of tasks should be taken into account when finding an approximate solution.

Thus, if a temperature field is determined on each elementary section of the fabric partition, the average temperature inside the fabric will be determined as the integral average over the entire surface of the fabric

$$T_{sr} = \frac{\sum_{i=1}^{m} T_i \cdot S_i}{S}, \qquad (11)$$

where: T_i – fabric temperature on i – split section;

S_i – area i split section;

S – total surface area of the fabric;

m – number of split sections.

Similarly to finding the average temperature, for example, the average coefficient of thermal conductivity of the fabric can be obtained.

IV. Numerical Solution of the Heat Conduction Equation and Comparison with Experimental Data

As an example of solving the problem of the one-dimensional heat equation (2), we consider a boundary

value problem for which, at the first stage, we make the following assumptions. We will analyze heat transfer through a thin single-layer shell of constant thickness, the initial temperature on the surface of the tissue adjacent to the body is constant and equal. The person is in a calm state and the additional heat production of the human body is not taken into account. Under the given conditions, the temperature will change only in the directions perpendicular to the surface of the fabric - the direction of the OX axis. Then, in the directions of the OУ and OZ axes, the temperature can be considered constant (Fig. 1)

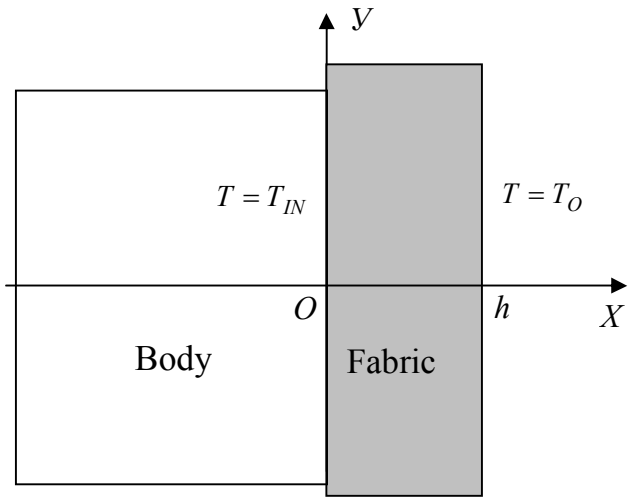


Fig. 1 – Geometry of smart fabric in a one-dimensional problem

We will also assume that the thermophysical characteristics of the fabric are not dependent on temperature. Then equation (2) is transformed to the form

$$c\rho\frac{\partial T}{\partial t} = \lambda\frac{\partial^2 T}{\partial x^2} - \frac{\alpha}{h}\left(T - T_O\right), \quad 0 < x < h. \tag{12}$$

The initial condition of the problem can be written in the form

$$t = 0, \quad T = T_O, \quad 0 \le x \le h. \tag{13}$$

The boundary conditions of the problem, which are conditions of the first kind, can be represented as

$$\begin{aligned} &x = 0, \quad T = T_{IN}, \quad t > 0; \\ &x = h, \quad T = T_O, \quad t > 0. \end{aligned} \tag{14}$$

Problem (12-14) will be solved by the finite difference method on a uniform grid. We divide the fabric in thickness into n intervals by points $x_k = \frac{kh}{n}, k = 0,1...n$. Determine the value of the temperature at the $j-$ node at the time $t = t_j = j\tau, j = 0,...,p$

$$T\left(x_k, t_j\right) = T_{kj},$$

where $\tau-$ time coordinate grid step, $j-$ time step number.

Replacing the differential operators in (12) by their finite-difference analogs, we obtain the following system of linear algebraic equations

$$\begin{aligned} &c\rho\frac{T_{k,j+1} - T_{k,j}}{\tau} = \lambda n^2\frac{T_{k+1,j+1} - 2T_{k,j+1} + T_{k-1,j+1}}{h^2} \\ &-\frac{\alpha}{h}\left(T_{k,j} - T_O\right), k = 1,...,n-1. \end{aligned} \tag{15}$$

This scheme for approximating the solution of problem (12-14) can be graphically represented as a four-point pattern of an implicit difference scheme

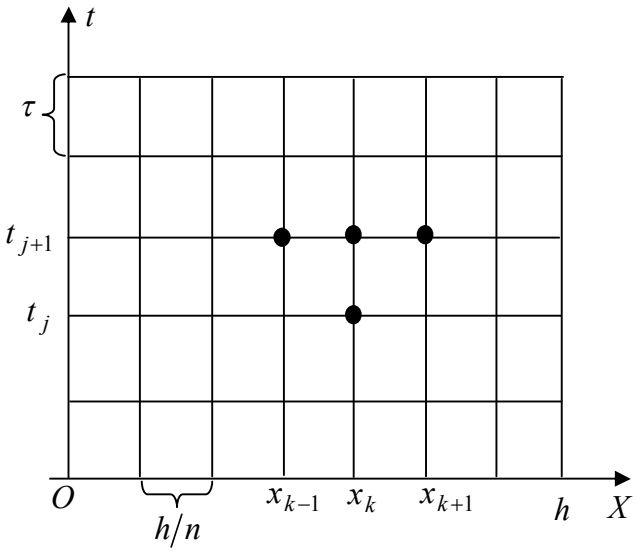


Fig. 2 – Four-point difference scheme template

The initial and boundary conditions (13, 14) imply the relations

$$\begin{aligned} &T_{k,0} = T_O, k = 0,...,n; \\ &T_{0,j} = T_{IN}, j = 1,...,p; \\ &T_{n,j} = T_O, j = 1,...,p. \end{aligned} \tag{16}$$

It can be shown that the implicit finite difference scheme (15, 16) is stable. For the numerical solution of problem (12, 14), it is also necessary to set its physical conditions. If the fabric is made of tencel, then for it

$$\lambda = 0{,}07\frac{W}{m\cdot K}, \rho = 120\frac{\text{kg}}{m^3}, c = 1400\frac{J}{\text{kg}\cdot K}.$$

Let us indicate the initial and boundary values of the problem

$$\begin{aligned} &h = 0{,}002m, T_O = 20^o C, \\ &T_{IN} = 36^o C, n = 10, \tau = 5s. \end{aligned}$$

The solution of the boundary value problem of thermal conductivity for the distribution of tissue temperature over thickness was carried out in the Maple environment after 60 seconds of heating. The temperature calculation results are shown in the table 1.

TABLE I. CALCULATION OF FABRIC TEMPERATURES BY THICKNESS ACCORDING TO THE DIFFERENCE SCHEME

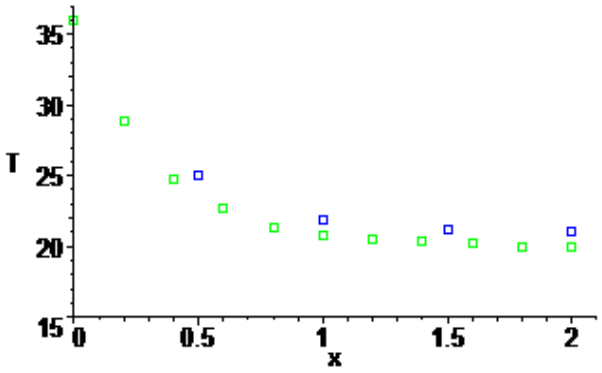
Point coordinates x_k, mm	0	0,2	0,4	0,6	0,8
Temperature values, °C	36	28,8	25	22,7	21,3
Point coordinates x_k, mm	1	1,2	1,4	1,6	1,8
Temperature values, °C	20,8	20,4	20,2	20,1	20

The temperatures of the tissue were also measured along the thickness with a mini thermometer Testo 0560 1110. The results of measurements of tissue temperature by thickness are shown in Table 2, as averaged over five independent and equally accurate measurements at each point.

TABLE II. THICKNESS AVERAGED FABRIC TEMPERATURES

Point coordinates x_k, mm	0	0,5	1	1,5	2
Temperature values, °C	36	24,9	21,9	21,1	21

A graphic comparison of the results of the numerical solution of the problem of heat conduction and temperature distribution according to the experiment is shown in the figure 3.



– temperature values according to the difference scheme;

– experimental temperature values.

Fig. 3 – Comparison of the numerical solution with the experimental results

The experimental temperatures turned out to be somewhat higher than those calculated using the difference scheme; however, the discrepancy between the calculated and experimental values does not exceed 7%.

V. CONCLUSIONS

The thermal conductivity of smart fabrics depends on a large number of factors: the type of material for a given site, the thickness and number of layers of the fabric, the nature of the fibers, the structural characteristics of the fabrics, porosity, temperature, humidity, etc.

As a result of the mathematical modeling of the thermal conductivity of fabrics, general approaches to determining the thermal conductivity of "smart" farbrics were described on the basis of the heat balance equation, from which, under the assumptions of the basic thermal conductivity coefficients in the elementary sections of the partition, the one-dimensional heat equation is consistent. Initial and boundary conditions were formulated for the heat equation, which simulate various conditions at the fabric boundary, as well as between its layers, and on the basis of which a unique based on finite difference schemes. The results of the experimental values of temperatures agree with the calculated values with an error of no more than 7%. It is shown that the average fabric temperature can be determined as an integral characteristic over the entire surface.

This paper is the basis for further determination of the local and integral characteristics of the thermal conductivity of smart fabrics.

REFERENCES

[1] X. Zhang and X. Tao, Smart textiles: Passive smart, *Textile Asia,* pp. 45-49, June 2001, Smart textiles: Very Smart, *Textile Asia,* pp. 35-37, August 2001.

[2] Textile institute, Smart Fibers, Fabrics and Clothing (Tao, X. Ed.), Florida: CRC Press, 2001.

[3] Smart Textiles: Smart Technology (http://www.ualberta.ca//smarttextiles), 4 October 2003.

[4] Smart-materials Overview, London UK (http://smarttextile.co.uk), 19 September 2006.

[5] K. Chapman, High Tech fabrics for smart garments, Concept 2 Consumer, pp. 15-19, September 2002.

[6] Mitsik, M.F., Bekmurzaev, L.A., Byrdina, M.V., Aleynikova, O.A., Kokhanenko, V.N. Description of the spatial shape surface of an air supported dynamic figure. Conference: 2019 IEEE East-West Design & Test Symposium (EWDTS). DOI: 10.1109/EWDTS.2019.8884390

[7] Byrdina, M.V., Mitsik, M.F., Bekmurzaev, L.A., Rubtsova, S.V. Surface visualization of flexible elastic shells. Conference: 2019 IEEE East-West Design & Test Symposium (EWDTS). DOI: 10.1109/EWDTS.2019.8884456

[8] Materials for clothing. Fabrics: textbook / B. A. Buzov, G. P. Rumyantseva. M.: ID "FORUM": INFRA-M, 2012. - 224 p. (Higher education).

[9] S.V. Lomov, D.S. Ivanov, G. Perie, I. Verpoest «Modelling 3D fabrics and 3D-reinforced composites: challenges and solution» 1st world conference on 3D fabrics, Manchester 9-11.04.2008.

[10] Yan J.Q., Zhang S.C., Kuzmichev V.E., Adolphe D.C. New database for improving virtual system "body-dress" In the collection: IOP Conference Series: Materials Science and Engineering 17, Shaping the Future of Textiles. Series "17th World Textile Conference, AUTEX 2017: Shaping the Future of Textiles - Fashion, Design and Garment Industry" 2017. C. 172029.

Developing a Multiple Testing Procedure in the D-Posterior Approach using the R Software Environment

1st Sergei Simushkin
Department of Mathematical Statistics
Kazan Federal University
Kazan, Russia
smshkn@gmail.com

2nd Elena Fedotova
Department of Mathematical Statistics
Kazan Federal University
Kazan, Russia
elvfdt@gmail.com

***Abstract*—The article provides a method for developing a statistical criterion that satisfies given restrictions on the average proportion of wrong decisions, similar to the concept of pFDR in multiple testing. The statistical criterion uses programming capabilities in the R environment. The advantages of this method are shown in comparison with the Benjamini-Hochberg procedure.**

***Keywords*—statistics, multiple testing, R programming language, d-posterior approach**

I. Introduction

The multiple testing problem always occurs when it's needed to test a large set of identical hypotheses simultaneously — e.g., when one would want to compare a large number of characteristics in a group of normal control subjects and experimental group. Common procedures designed to control the type I error can't be applied in this case since they don't guarantee that a total (in one sense or another) error rate will be small enough. There are different corrections often used to solve this problem, for example, the Bonferroni correction, designed to control the family-wise error rate (FWER) [1]. Though this correction is easy to understand and implement it significantly reduces the power of experiment when we're testing more than a few hypotheses.

Sometimes, instead, it's more reliable to use the Benjamini–Hochberg procedure that allows to set a limit on the value of FDR — expected proportion of false positives:

$$FDR = E\left(\frac{V}{R \vee 1}\right) = E\left(\frac{V}{R} \;\middle|\; R > 0\right)\mathbf{P}(R > 0),$$

where V — the number of false discoveries (positives), R — the number of rejected null hypotheses. The Benjamini–Hochberg procedure is a step-up procedure that based on adjusting the p-value [2].

For all its merits, this procedure suffers from a number of significant disadvantages. If the actual proportion of true null hypotheses is close to 100%, the Benjamini–Hochberg procedure may lose the guarantee property. In the opposite case, when this proportion is small, the procedure becomes too conservative and decreases the power of the experiment [3]. Moreover, the Benjamini–Hochberg procedure can't be used to control the FNR (false negative rate) at all.

As an alternative approach Storey [4] proposed using so-called pFDR (positive false discovery rate) to solve the multiple testing problem. The pFDR is defined as follows

$$pFDR = E\left(\frac{V}{R} \;\middle|\; R > 0\right).$$

In the same article Storey also consider solving the multiple testing problem within the framework of the Bayesian paradigm. In this case it's assumed that the control parameter in each testing is the realization of some random variable. It's remarkable that if all hypotheses are independent, then the concept of the pFDR is identical to the value of the type I d-risk, which was introduced in [5]. The approach in which the risk of a statistical procedure is calculated as the conditional average of losses among experiments that ended up with the same decision was called d-posterior in [5].

Using methods of the d-posterior approach requires knowledge of the entire structure of the probabilistic model — both the distribution of the test statistic and the distribution of the control parameter. If there is no need to build an optimal statistical procedure, but to use a specific test statistic, for example p-value, then some approximations of each part of this model can be proposed. The main issue of this approach is the need to make some time-consuming calculations. To solve this issue we used the R software environment.

In this article, the d-posterior approach is applied to the problem of identifying genes that have different expression level in the group of patients with cancer compared to their expression level in the normal control group. All computational procedures are implemented in R programming language. Statistical data are taken from the open source [6].

The work was funded according to the development program of Scientific and Educational Mathematical Center of Volga-region Federal District, project N 075-02-2020-1478

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

II. Statistical Model

Without going into the depth of the d-posterior approach (see, for example, [7]), we consider the specific problem of comparing the level of gene expression in two groups. The data represent the expression level of $M = 6033$ genes measured in each of the $n = 50$ patients in the control group and $k = 52$ patients in the experimental group suffering from cancer. The data are partly shown in the Table I. Each row of the table represents the expression of a particular gene taken from normal control subjects (indicated by x) and subjects with cancer (indicated by y).

TABLE I
THE SAMPLE DATA WITH GENE EXPRESSION LEVELS

	x_1	x_2	x_3	...	y_1	y_2	y_3	...
gene 1	-0.93	-0.75	-0.55	...	-1.09	-0.58	-1.09	...
gene 2	-0.84	-0.85	-0.85	...	-0.83	0.25	-0.83	...
gene 3	0.06	0.10	-0.003	...	1.23	0.11	4.04	...
gene 4	-0.36	2.42	-0.12	...	0.79	-0.12	-0.35	...
gene 5	-1.12	0.18	0.84	...	0.69	0.94	-1.08	...
...	...	...	...	...	...	...	...	...

Let $(x_1, \dots, x_n), (y_1, \dots, y_k)$ be the observed values of the expression level of some (say, i-th) gene in the control group and in the experimental group (with cancer), respectively. To compare the expected levels of expression in groups, we calculate Welch's statistic

$$t_i = \frac{\overline{y} - \overline{x}}{\sqrt{s_x^2/n + s_y^2/k}},$$

where $\overline{x}, \overline{y}$ are sample means, s_x^2, s_y^2 are sample variances of the control group and the experimental group, respectively.

It is known that if the difference between the expected expression levels of some gene in two groups is θ, then even with small sample sizes n, k the distribution of Welch's statistic T is well approximated by the normal distribution with mean $c\,\theta$ and variance 1, where the value of constant c depends only on the sample sizes n, k. The way to calculate the p-value depends on the considered hypotheses. If the null hypothesis $\mathbf{H}_0 : \theta \le 0$ is tested with the alternative $\mathbf{H}_1 : \theta > 0$ (the disease leads to an increase in the expression level), then the p-value is calculated as

$$\Pi(t) = 1 - \Phi(t),$$

where Φ is the cumulative distribution function of the standard normal distribution, $t = t_i$ is the value of Welch's statistic obtained by analyzing the i-th gene. It is well known that when $\theta = 0$ the distribution of the p-value, taking into account the normal approximation of the distribution of T, can be approximated by a standard uniform distribution. When $\theta \neq 0$ the following statement can be proved:

Lemma 1:

I. If the statistic T is normally distributed with mean θ and variance 1, then the cumulative distribution function

$$\mathbf{P}_\theta\big(\Pi(T) < x\big) = \mathbf{P}_\theta\big(1 - \Phi(T) < x\big),\ x \in [0,1],$$

is concave when $\theta > 0$ (i.e. inside the alternative) and convex when $\theta < 0$ (inside the null).

II. The cumulative distribution function of the beta distribution with shape parameters (a, b) is concave when $a \le 1, b \ge 1$ and convex when $a \ge 1, b \le 1$.

It was shown that every distribution defined on the interval $[0, 1]$ can be approximated by the beta distribution [8]. Parameters of the beta approximation should be chosen considering convexity properties of the cumulative distribution function of the p-value. Let the probability density function of the beta distribution be equal to $Cx^{a-1}(1-x)^{b-1},\ x \in [0,1]$. Then for convex distributions it's relevant to choose parameters $a \ge 1, b \le 1$ and for concave distributions — $a \le 1, b \ge 1$.

In the article, we consider a simplified model when it is assumed that the difference in gene expression levels takes only three possible values: $\theta = 0$ — gene expression does not differ in groups, $\theta = \theta_1 < 0$ — gene expression decreases in the experimental group, $\theta = \theta_2 > 0$ — gene expression increases in the experimental group.

Thus, the distribution of the p-value can be represented as a mixture of three components: with the probability π_0, the statistic $\Pi(T)$ has a standard uniform distribution (which is equivalent to the beta distribution with shape parameters $a = b = 1$), with the probability π_1 the statistic $\Pi(T)$ has the beta distribution with parameters $a > 1, b < 1$, finally, with the probability π_2 the statistic $\Pi(T)$ has the beta distribution with parameters $a < 1, b > 1$. So the cumulative distribution function of the statistic $\Pi(T)$ with this representation is as follows

$$F(t) = \mathbf{P}\big(\Pi(T) < t\big) = (1 - \pi_1 - \pi_2)\mathtt{B}(t\,;\,1,1) + \\ + \pi_1 \mathtt{B}(t\,;\,a_1, b_1) + \pi_2 \mathtt{B}(t\,;\,a_2, b_2), \quad (1)$$

where $0 \le \pi_1, \pi_2 \le 1,\ \pi_1 + \pi_2 \le 1,\ a_1 > 1, b_1 < 1,\ a_2 < 1, b_2 > 1$, $\mathtt{B}(t\,;\,a,b)$ — the cumulative distribution function of beta distribution with shape parameters a, b.

Similar mixture model but with two components of beta was already used to approximate p-value's distribution (see, for example, [9] and [10]). Since we wanted to study the case when it's needed to test a one-sided alternative, we have chosen the three-component model. In most cases, when p-values was obtained from a test with one-sided alternative, its histogram has two peaks — near 0 and 1. So the three-component mixture model makes it possible to simulate the peaks as well as the uniform distribution in the middle of the histogram.

III. Parameters Estimation

The model (1) has six unknown parameters. Since the analyzed sample is large (more than 6000 observations), the parameters can be estimated quite accurately. We have applied the maximum likelihood method for this data. All calculations were carried out in R software environment. The following is a detailed description of the algorithm:

1) Welch's statistics were calculated for each gene: $t_1, \dots, t_M$;

2) the sample of p-values $p^{(M)} = (p_1, \ldots, p_M)$ was obtained as follows

$$p_i = \Pi(t_i) = \mathbf{P}\left(T > t_i \,|\, \theta_i \le 0\right) = 1 - \Phi(t_i),$$

where $i = 1, \ldots, M$.

3) the negative log-likelihood of the model (1) was determined as follows:

$$L(\pi_1, \pi_2, a_1, b_1, a_2, b_2 \,|\, p^{(M)}) = \\ = -\sum_{i=1}^{M} \log[\pi_0 + \pi_1 b(p_i\,;\, a_1, b_1) + \\ + \pi_2 b(p_i\,;\, a_2, b_2)],$$

where $\pi_0 = 1 - \pi_1 - \pi_2, b(\,\cdot\,;\, a, b)$ — the probability density function of the beta distribution with shape parameters $a, b > 0$.

4) the minimum of the function L was found with function `mle2` from the package `bbmle` [11]. To keep the convexity properties of the cumulative distribution function of the p-value we set the following constraints on the optimization parameters:

$$a_1 > 1, b_1 < 1,$$
$$a_2 < 1, b_2 > 1.$$

We also added obvious constraints on the probabilities

$$0 \le \pi_1, \pi_2 \le 1,$$
$$\pi_1 + \pi_2 \le 1.$$

To minimize the function with constraints the package `bbmle` has the L-BFGS-B algorithm.

As a result of the calculations, the following parameter estimates were obtained:

$$\pi_0 = 1 - \pi_1 - \pi_2 = 0.9004,$$
$$\pi_1 = 0.0497,\ a_1 = 2.9565,\ b_1 = 0.3905, \qquad (2)$$
$$\pi_2 = 0.0499,\ a_2 = 0.3508,\ b_2 = 2.5053.$$

According to the estimates obtained, we can say that it's a high chance (90%) for a gene not to be responsible for the disease. This is a common situation in microarray studies that only a small proportion of genes is interesting.

Data histogram and fitted density function with parameter estimates (2) of the model (1) are represented on the Figure 1. It seems that the provided beta mixture distribution is quite a good approximation to the sample p-values.

IV. Guarantee Criterion

The null hypothesis is rejected if the p-value is $\Pi(t) < c$, where the constant c must satisfy the condition for the d-posterior probability of the type I error:

$$\mathcal{R}_1(c) = \mathbf{P}\big(\theta \le 0 \,|\, \Pi(T) < c\big) = \\ = \frac{\pi_0 \mathtt{B}(c\,;\, 1, 1) + \pi_1 \mathtt{B}(c\,;\, a_1, b_1)}{F(c)} \ \le\ \beta_1, \quad (3)$$

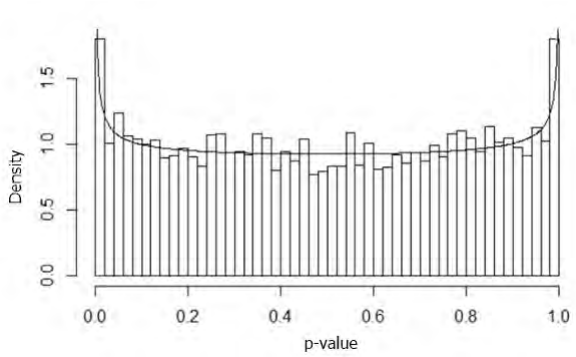


Fig. 1. Histogram of sample data with expression levels of normal control subjects and cancer patients. Curve is fitted density function from the beta mixture distribution with parameter estimates (2). It seems that the beta mixture model is a good approximation to the distribution of real sample p-values.

i.e. the conditional probability that gene expression doesn't actually increase in the experimental group, but we decided otherwise, is less than the specified level of β_1.

Note that in a similar way one can develop a procedure with a restriction on the d-posterior probability of the type II error:

$$\mathcal{R}_0(c) = \mathbf{P}\big(\theta > 0 \,|\, \Pi(T) \ge c\big) = \frac{\pi_2(1 - \mathtt{B}(c\,;\, a_2, b_2))}{1 - F(c)} \ \le\ \beta_0.$$

In order to increase the power, the constant c should be chosen so that an equal sign is reached in (3). In R environment, you can use the `uniroot()` procedure for this.

The critical constant for the model with the parameters (1) and the level $\beta_1 = 0.10$ turned out to be equal to $c = 0.00073$. Among the $M = 6033$ genes, only in 45 cases the p-value of the Welch's test was less than this constant. Applying the Benjamini-Hochberg procedure to the same data also led to 45 discoveries.

In connection with the latter, we note that this situation rarely occurs. As further studies have shown, the Benjamini-Hochberg procedure most often reveals fewer discoveries than the d-posterior procedure. In addition, the Benjamini-Hochberg procedure does not provide any information about the probability of the type II error or FNR (false negative rate).

On the contrary, in the framework of the d-posterior approach, one can solve the equation $\mathcal{R}_0(c) = \beta_0$ and build a procedure with restrictions on the proportion of false negatives. For example, solving the equation $\mathcal{R}_0(c) = \beta_0$ with $\beta_0 = 0.01$ gives the constant $c = 0.3031$. Note that the probability of the type II error is $\mathcal{R}_0(c) \le \pi_2$. Therefore, if the restrictions on the d-posterior probability of the type II error are not too strict ($\beta_0 > \pi_2$), then we should decide that the expression level of all genes does not increase with the disease.

V. Accuracy of the Fitted Model

Since we estimated the model parameters to construct statistical inference procedures, it becomes necessary to know how much influence the chosen estimation method has on the properties of the decision-making procedure. So we used a

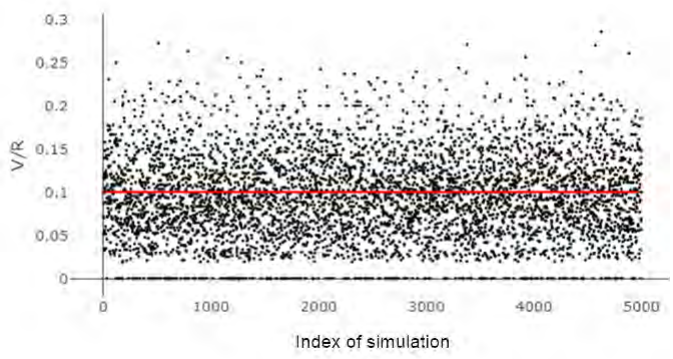


Fig. 2. Scatter plot of V/R obtained from simulation results. It seems that the proportion of false positives is centered around the given restriction on the type I d-risk, which represented by the solid line

stochastic simulation to examine the accuracy of our procedure. The following calculations were performed 5000 times:

1) A sample is generated from the multinomial distribution with the support $\{\theta = 0, \theta < 0, \theta > 0\}$ and the according probabilities $\{\pi_0, \pi_1, \pi_2\}$ using the function `sample()`.
2) For each value in the sample, a p-value is generated following the corresponding distribution:
   a) $\theta = 0 \Rightarrow B(1, 1)$;
   b) $\theta < 0 \Rightarrow B(a_1, b_1)$;
   c) $\theta > 0 \Rightarrow B(a_2, b_2)$.
3) The described d-posterior procedure is carried out with $\beta_1 = 0.1$, then the number of false positives V and the number of rejected null hypotheses R are calculated.
4) For comparison, the Benjamini-Hochberg procedure at level 0.1 is carried out and the values V_{bh} and R_{bh} are calculated similarly.

The results of simulation are shown in Figure 2. It seems that the proportion of false positives V/R is centered around the restriction on the type I d-risk which is equal to 0.1 in our case.

According to the results of the stochastic simulation, the estimation of the type I d-risk and FDR are as follows:

$$\mathcal{R}_1^* = \frac{AVG(V)}{AVG(R)} \approx 0.098,$$

$$FDR^* = AVG\left(\frac{V}{R \vee 1}\right) \approx 0.095,$$

$$FDR_{BH}^* = AVG\left(\frac{V_{bh}}{R_{bh} \vee 1}\right) \approx 0.08873,$$

where $AVG(\cdot)$ is the sample mean.

As you can see the proposed d-posterior method actually keeps the value of the type I d-risk at a given level. Moreover, in our case, we got the value of FDR at the same level. As for the Benjamini-Hochberg procedure, we can confirm that as it was expected the procedure controls the value of FDR at the lower level. Since the BH procedure actually controls the FDR at level $\frac{n_0}{M}\beta_1$, where n_0 — the number of true null hypotheses, $\beta_1 = 0.1$, the estimation of the proportion $\frac{n_0}{M}$ is around 0.88, which is approximately the maximum likelihood estimation of π_0.

VI. Conclusion

Thus the use of the beta distribution in multiple testing problems within the framework of the d-posterior approach makes it possible to develop statistical criteria with restrictions on the probability of errors among experiments ending with the rejection of the null hypothesis, as well as among experiments ending with the acceptance of the null hypothesis, which was the main goal of this research.

R programming language procedures that was developed to fit the beta mixture model to the data showed good result and allowed us to perform complex numerical experiments to analyze and confirm the claimed warranty properties of the proposed criteria.

This paper also provides detailed algorithm to implement the method in the R software environment and shows that this method actually allow us control the type I d-risk on the given level. The novelty of the research is that the proposed method allows one to correctly perform large-scale statistical experiments, that is really common in our big data era.

Though the procedure was used in gene expression study, it can be applied to any multiple testing problem if only the test hypotheses are identical and independent, the control parameter can be interpreted as a random variable and the test statistic can be approximated by the normal distribution, which is satisfied in most cases. There will be further study about the applicability of this method in the cases of dependence and non-normality of the test statistic.

References

[1] Hochberg Y., Tamhane A.C. "Multiple Comparison Procedures", New York : Wiley, 1987
[2] Benjamini Y., Hochberg Y. "Controlling the false discovery rate: a practical and powerful approach to multiple testing", Journal of the Royal statistical society: series B (Methodological), 1995, vol. 57, no. 1, pp. 289-300.
[3] Zaykin D. V., Young S. S., Westfall P. H. "Using the false discovery rate approach in the genetic dissection of complex traits: a response to Weller et al Genetics", 2000, vol. 154, no. 4, pp. 1917-1918.
[4] Storey J. D. "The positive false discovery rate: A Bayesian interpretation and the q-value", Annals of Statistics, 2003, vol. 31, pp. 2013-2035.
[5] Simushkin S.V. "Confidence Bounds and Narrowest Reliable Intervals in D-Posterior Approach", Lobachevskii Journal of Mathematics, 2018, vol.39, no.3, pp.388-397.
[6] Stanford University, School of Humanities & Sciences, Statistics, Bradley Efron personal page, datasets and programs. [Online]. Available: http:statweb.stanford.edu/ ckirby/brad/LSI/datasets-and-programs/data/
[7] Simushkin D.S., Simushkin S.V., Volodin I.N. "D-guaranteed discrimination of statistical hpotheses: review of results and unsolved probems", Journal of Math. Sci., 2018, vol. 228, no. 5, pp. 543-565.
[8] Parker R.A., Rothenberg R.B. "Identifying important results from multiple statistical tests", Statistics in Medicine, 1988, vol.7, pp. 1031-1043
[9] Allison D.B., Gadbury G.L., Fernandez J.R., Lee C.K., Prolla T.A., Weindruck R. "A mixture model approach for the analysis of microarray gene expression data", Computational Statistics & Data Analysis, 2002, vol. 39, pp. 1–20.
[10] Yu C., Zelterman D. "A parametric model to estimate the proportion from true null using a distribution for p-values", Computational Statistics & Data Analysis, 2017, vol. 114, pp. 105–118.
[11] Bolker B. (2020) The bbmle package. [Online]. Available: https://cran.r-project.org/web/packages/bbmle/bbmle.pdf

Relationship Between Base Frequency of the Koch-Type Wire Dipole and Various Dimensions

Ilya Pershin
Institute of Computational Mathematics and Information Technologies
Kazan Federal University
Kazan, Russia
muzclubs@gmail.com

Dmitrii Tumakov
Institute of Computational Mathematics and Information Technologies
Kazan Federal University
Kazan, Russia
dtumakov@kpfu.ru

***Abstract*—A dipole wire antenna of the Koch type is considered. The antenna consists of a wire dipole with symmetrical arms with respect to the feed point with the geometry similar to the Koch prefractal. The curves forming the arms differ from the classical Koch fractal only by the position of the central vertex. The work's goal is to establish the dependence of the base frequency on the dimension of the curve forming the antenna arm. Various dimensions as characteristics of the curve are considered. The dimensions are Minkowski dimension, information dimension, correlation dimension and Higuchi fractal dimension. The algorithm to calculate the Higuchi dimension for our curves is adapted. Also, algorithms for calculating the other dimensions are described. Relationships between the base frequency of the Koch-type wire dipole and the dimensions are explored. The correlation analysis for the first three Koch-type prefractals is carried out. The values of all correlation coefficients between the base frequency and the considered dimensions are given in the tables. It is concluded that for the second and third iterations, the best correlation is a correlation between the base frequency and the Higuchi dimension. The optimal one-parameter regression models for the base frequency in the case of the second and third iterations are constructed. The obtained regression model for the second iteration approximates the frequency values with an error of 1.17%. The model for the third iteration approximates the frequency values with an error of 1.46%.**

***Index Terms*—Koch-type wire dipole, base frequency, curve dimension, correlation analysis**

I. Introduction

Wire antennas of various types are widely used in modern telecommunication systems [1]. Nowadays, antennas with a complex geometry are the most promising devices [2]. In practice, various forms of broken balanced dipoles are used [3], [4]. However, the most widespread method to minimize antenna size is fractalization [5]–[11]. At the moment, the most studied fractal antennas are those built on the basis of the Koch fractal [12]. First of all, it is the Koch monopole [13], the Koch dipole [14] and Koch log periodical antenna [15]. In addition, dipoles of the Koch type are used [16].

For the simplest half-wave dipole, especially in the case of the base frequency, the interconnections between its various parameters are well known [17]. The relationship between the base frequency and the reflector geometry was also investigated by the authors for Koch-type dipoles [16], [18]. Such a relationship can be used for the designing of antennas [19].

The authors previously used this approach to model a Koch-type dipole of the first iteration for Wi-Fi applications [20]. Also, regression models for the base frequency are used for monopole antennas [21]. Such high-precision models can significantly accelerate the process of designing antennas [22].

The base frequency (wavelength) depends directly on the length of the wire forming a half-wave dipole. However, this dependence has dispersion for arms of a complex geometry, and for Koch-type dipole this effect is also observed.

Therefore, to increase the efficiency of the antenna design algorithm, it is desirable to have models approximating the frequency with high accuracy. The purpose of the present work is to establish the dependencies between the base frequency and fractal dimensions, as well as to obtain the most accurate regression model.

II. Problem statement

Let us consider a symmetrical wire dipole with power supply located in the middle of the antenna. The dipole has a geometry of arms similar to the Koch-type prefractal of the first, second and third iterations [23]. Initiating curves for the Koch-type fractal differ from the initiating curves of the classical Koch fractal only by the position of the central vertex.

A Koch-type fractal is not a classical fractal; it received its name due to its similarity to the Koch fractal [24]. The first iterations of these fractals differ only in the position of the central vertices (see Fig. 1). When constructing iterations of Koch-type fractals, a fractal interpolation algorithm is used [25], [26]. A description of this algorithm is given in detail below.

Let

$$K_1 = \{(t_i, x_i, y_i) \in [0,1] \times R \times R \,| \\ 0 = t_0 < t_1 < ... < t_n = 1\}$$

be some interpolation points for $i = 1..n$. For the Koch-type fractal, the number $n = 4$ is chosen. The values $x_0 = 0, x_1 = 1/3, x_3 = 2/3, x_4 = 1$ and $y_0 = y_1 = y_3 = y_4 = 0$

978-1-7281-9899-6/20/$31.00 2020 IEEE

at arbitrary x_2 and $y_2 > 0$ form interpolation points for a family of Koch-type curves. Also, for antenna applications, the restriction $0 < x_2 < x_4$ is imposed. Note that in the case $x_2 = 1/2$, $y_2 = 1/\sqrt{12}$, the first iteration of the classical Koch curve is obtained.

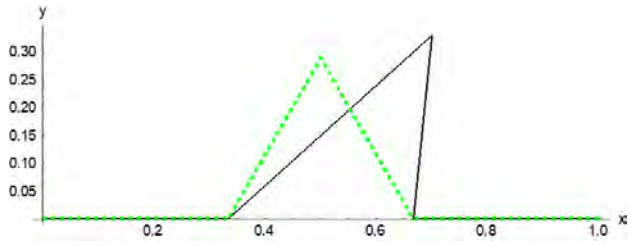


Fig. 1. First iteration. Black solid line is for the Koch-type fractal; green dashed line is for the classical Koch fractal

For all i, the affine transformation is introduced according to the following rule: $A_i : R^3 \to R^3$,

$$A_i \begin{pmatrix} t_i \\ x_i \\ y_i \end{pmatrix} := \begin{pmatrix} a_i & 0 & 0 \\ c_{i1} & D_{i1} & D_{i2} \\ c_{i2} & D_{i3} & D_{i4} \end{pmatrix} \begin{pmatrix} t_i \\ x_i \\ y_i \end{pmatrix} + \begin{pmatrix} e_i \\ d_{i1} \\ d_{i2} \end{pmatrix}. \quad (1)$$

Here, the matrices D_{ij} have the form:

$$D_{i1} = D_{i4} = \begin{pmatrix} 0.333 & 0 \\ 0 & 0.333 \end{pmatrix},$$

$$D_{i2} = \begin{pmatrix} 0.167 & -0.289 \\ 0.289 & 0.167 \end{pmatrix},$$

$$D_{i3} = \begin{pmatrix} 0.167 & 0.289 \\ -0.289 & 0.167 \end{pmatrix}. \quad (2)$$

Next, we require that all i satisfy the following conditions:

$$A_i \begin{pmatrix} t_0 \\ x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} t_{i-1} \\ x_{i-1} \\ y_{i-1} \end{pmatrix}, \quad A_i \begin{pmatrix} t_4 \\ x_4 \\ y_4 \end{pmatrix} = \begin{pmatrix} t_4 \\ x_4 \\ y_4 \end{pmatrix}. \quad (3)$$

Then, the following is obtained:

$$a_i = t_i - t_{i-1}, e_i = t_{i-1}, \quad (4)$$

$$\begin{pmatrix} c_{i1} \\ c_{i2} \end{pmatrix} = \begin{pmatrix} x_i - x_{i-1} \\ y_i - y_{i-1} \end{pmatrix} - \begin{pmatrix} D_{i1} & D_{i2} \\ D_{i3} & D_{i4} \end{pmatrix} \begin{pmatrix} x_4 - x_0 \\ y_4 - y_0 \end{pmatrix},$$

$$\begin{pmatrix} f_{i1} \\ f_{i2} \end{pmatrix} = \begin{pmatrix} x_{i-1} \\ y_{i-1} \end{pmatrix} - \begin{pmatrix} D_{i1} & D_{i2} \\ D_{i3} & D_{i4} \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}. \quad (5)$$

With this definition of the operators A_i, the straight line segment connecting the points x_0 and x_4 merges into a polyline, connecting the interpolation points in a consecutive manner.

Thus, a set of points for the second and third iterations is obtained as follows:

$$K_2 = \bigcup_{i=1}^{n} A_i(K_1), \quad K_3 = \bigcup_{i=1}^{n} A_i(K_2). \quad (6)$$

The graph of the obtained Koch-type prefractals is presented in Fig. 2.

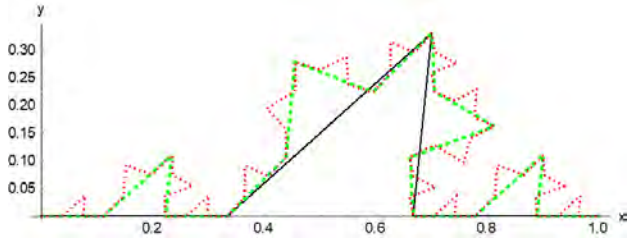


Fig. 2. First (black solid line), second (green dashed line), and third (red dotted line) iterations of the Koch-type fractal. Coordinates of the central vertex: $x_2 = 0.7$, $y_2 = 0.33$

For the antennas formed by the first prefractal iteration, we will vary the coordinates of the central vertex at the same interval for x and y changing from 0.25 cm to 7.46 cm with the step of 1.025 mm. For the antennas with the second iteration: x varies from 7.5 mm to 6.45 cm with the step of 3 mm; x varies from 3 cm to 6.5625 cm with the step of 2.0625 mm. For the third iteration, x changes from 2.625 mm to 5.625 cm in increments of 2.625 mm; y changes from 2.625 cm to 6.0 cm in increments of 2.625 mm. Thus, we obtain 225 antennas for the first iteration, 361 for the second iteration and 399 for the third iteration. Let us also exclude the antennas that have self-intersections.

By the method of paired correlation analysis we investigate this family of Koch-type dipoles, establish the dependence of the base frequency f_1 on the dimension of the curve forming the antenna arms. Let us also consider Minkowski D_M, information D_I, correlation D_C and Higuchi D_H dimensions.

On the basis of the correlation analysis, we will construct regression models of the base frequency f_1 of the dipole from the dimensions. All calculations will be performed for the antenna with wire diameter $d = 2$ mm.

Minkowski dimension D_M, information dimension D_I and correlation dimension D_C will be calculated for the curve of 4 pixels width, located on the image with the size of 5000 by 5000 pixels, starting with the grid step of 5 pixels.

III. Dimensions

We define the various dimensions in the section.

A. Minkowski dimension

Firstly, let us consider the Minkowski dimension D_M [27]. This dimension shows the degree of self-similarity of the fractal. We describe the algorithm of the Minkowski dimension calculation.

Let E be any non-empty limited set at $\mathbb{R}^n$. Then the Minkowski dimension is defined as

$$D_M = \lim_{\varepsilon \to 0} \frac{\ln N(E, \varepsilon)}{-\ln \varepsilon}, \quad (7)$$

where $N(E, \varepsilon)$ is the minimum number of sets of diameter equal to ε that can cover the initial set E.

In practice, the box-counting dimension algorithm is used to calculate the Minkowski dimension. The box-counting function of E is a function $N_B(E, \varepsilon) \to \mathbb{N} \cup \{0\}$, where $N_B(E, \varepsilon)$

denotes the minimum number of balls B with a diameter ε that can cover the original set E:

$$D_{bc} = \frac{\ln N_B(E,\varepsilon)}{-\ln \varepsilon}.$$

The Minkowski dimension value is defined as the tilt angle tangent of the dependency graph of $\ln N_B(E,\varepsilon)$ on $\ln(1/\varepsilon)$. Also note that in the two-dimensional case, when the source area is a rectangle, it is divided into squares with the size of the side equal to ε.

B. Information dimension

Let us consider the information dimension D_I. This dimension shows the growth rate of information entropy and is determined as follows [28]:

$$D_I = \lim_{\varepsilon \to 0} \frac{S(\varepsilon)}{\ln \varepsilon}, \tag{8}$$

where

$$S(\varepsilon) = -\sum_{i=1}^{N(\varepsilon)} p_i(\varepsilon) \ln p_i(\varepsilon).$$

By p_i, we mean the probability of a point hitting the ball B_i of diameter ε or, in other words,

$$p_i(\varepsilon) = \lim_{N \to \infty} \frac{n_i(\varepsilon)}{N},$$

where n_i is the number of points in the ball B_i, and N is the total number of points in the set E.

The information dimension value is defined as the tilt angle tangent of the dependency graph of $S(\varepsilon)$ on $\ln \varepsilon$.

C. Correlation dimension

Let us consider the correlation dimension D_C. It is defined as follows [29]:

$$D_C = \lim_{\varepsilon \to 0} \frac{\ln I(\varepsilon)}{\ln \varepsilon}, \tag{9}$$

where

$$I(\varepsilon) = \sum_{i=1}^{N(\varepsilon)} p_i^2(\varepsilon).$$

The value $p_i^2(\varepsilon)$ represents the probability of two points hitting the ball B_i with a diameter of ε. Summing up $p_i^2(\varepsilon)$ for all occupied balls B_i, we obtain the probability that two arbitrarily chosen points from the set E lie within one ball with the diameter ε.

The correlation dimension value is defined as the tilt angle tangent of the dependency graph of $\ln I(\varepsilon)$ on $\ln \varepsilon$.

D. Higuchi dimension

Now let us consider the Higuchi dimension D_H [30]. This dimension is used to analyze time series. Such series are characterized by a uniform time step Δt. Let us bring our curve to "that kind of appearance". In order to do this, we divide the curve into equal lengths of Δt. For time series values, let us take the values of the curve $y(t)$ by obtaining a set of discrete data: $y(1), y(2), \ldots, y(N)$.

Let us consider the following sequences at $m = 0, \ldots, d$:

$$S_m(d) = y(m), y(m+d), y(m+2d), \ldots, y\left(m + \left\lfloor \frac{N-m}{d} \right\rfloor d\right).$$

Here $\lfloor x \rfloor$ is an entire part of x. For each sequence $S_m(d)$, we calculate the length

$$L_m(d) = \frac{N-1}{\left\lfloor \frac{N-M}{d} \right\rfloor d} \sum_{i=1}^{\left\lfloor \frac{N-M}{d} \right\rfloor} |y(m+id) - (m+(i-1)d)|,$$

where m and d are integers. Let us average lengths $L_m(d)$ and obtain

$$L(d) = \frac{1}{d} \sum_{m=0}^{d} L_m(d).$$

The Higuchi dimension value is defined as the tilt angle tangent of the dependency graph of $\ln L(d)$ on $\ln d$:

$$D_H = \frac{\ln L(d)}{\ln d}, \quad d \gg 1. \tag{10}$$

For the calculations, we choose the interval d from 3 to 40 with the step of 1.

IV. Correlation and regression analysis

Let us carry out the correlation analysis for the base frequency and dimensions of the dipole arm curves in case of the first three iterations of the Koch curve. Let us consider the first iteration. To calculate the dimensions (7)–(9), let us divide the initial area covering the curve of Koch type into squares with the length of the side ε. Moreover, ε is selected such that it completely covers the source area without crossing. Values of D_H are calculated by formula (10) with step $\Delta t = 0.01$. In Table I, we present values of correlation coefficients.

TABLE I
Correlation coefficients for the first iteration

	f_1	D_M	D_C	D_I	D_H
f_1	1	-0.953	-0.967	-0.962	-0.835
D_M	-0.953	1	0.976	0.964	0.890
D_C	-0.967	0.976	1	0.999	0.921
D_I	-0.962	0.964	0.999	1	0.926
D_H	-0.835	0.890	0.921	0.926	1

You may note that the correlation between f_1 and D_H is worse than that between other dimensions. Meanwhile, the other dimensions have approximately the same correlation (a little more than 0.95) with the base frequency. Next, we consider the values of the correlation coefficients for the second iteration of the Koch-type curve (see Table II). In this case, the curve splitting for the Higuchi dimension is done with the step $\Delta t = 0.02$.

For the second iteration, the D_H to f_1 correlation is the strongest; it exceeds 0.99. Let us build a one-parameter regression model for the base frequency:

$$\hat{f}_1 = 1086.1 - 1068.57 D_H, \tag{11}$$

TABLE II
CORRELATION COEFFICIENTS FOR THE SECOND ITERATION

	f_1	D_M	D_C	D_I	D_H
f_1	1	-0.932	-0.974	-0.963	-0.993
D_M	-0.932	1	0.940	0.902	0.927
D_C	-0.974	0.940	1	0.994	0.970
D_I	-0.963	0.902	0.994	1	0.960
D_H	-0.993	0.927	0.970	0.960	1

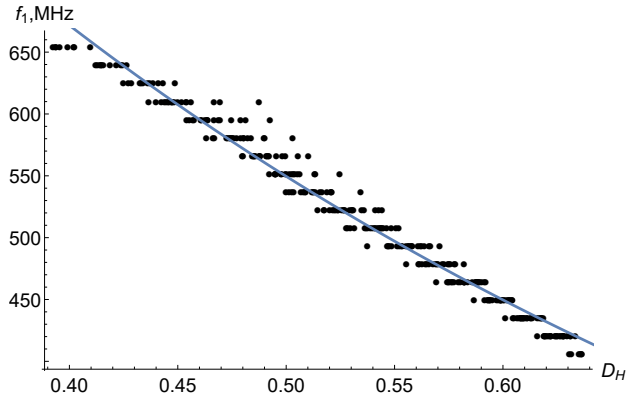


Fig. 3. Ratios of D_H to f_1 for the second iteration

where $\hat{f}_1$ is measured in MHz. Formula (11) approximates the base frequency with the relative error $\delta \approx 1.17\%$. The regression line is shown in Fig. 1.

Now let us study the correlation between the base frequency and dimensions for the third iteration. Let us calculate D_H with step $\Delta t = 0.02$. Results of calculations of all correlation dependencies are presented in Table III.

TABLE III
CORRELATION COEFFICIENTS FOR THE THIRD ITERATION

	f_1	D_M	D_C	D_I	D_H
f_1	1	-0.950	-0.962	-0.961	-0.980
D_M	-0.950	1	0.989	0.981	0.943
D_C	-0.962	0.989	1	0.998	0.959
D_I	-0.961	0.981	0.998	1	0.959
D_H	-0.980	0.943	0.959	0.959	1

The results for the Minkowski dimension (7) are shown in Fig. 4. Also in Fig. 4 the regression curve is shown, which is defined by the following formula:

$$\hat{f}_1 = -56075.8\, D_M^2 + 146468\, D_M - 94790.4. \qquad (12)$$

The formula (12) has the relative error about 4.77%.

The correlation field for the information dimension (8) is presented in Fig. 5. It is easy to see that the points in this case are scattered less than those in Fig. 4. The regression formula

$$\hat{f}_1 = -55824.9\, D_I^2 + 160201\, D_I - 114105 \qquad (13)$$

has a smaller error: $\delta \approx 3.41\%$.

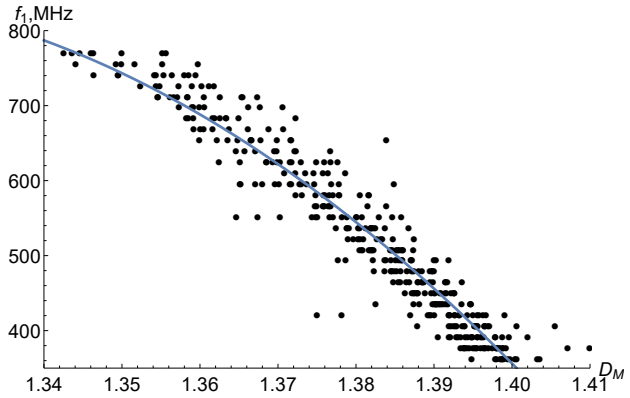


Fig. 4. Ratios of D_M to f_1 for the third iteration

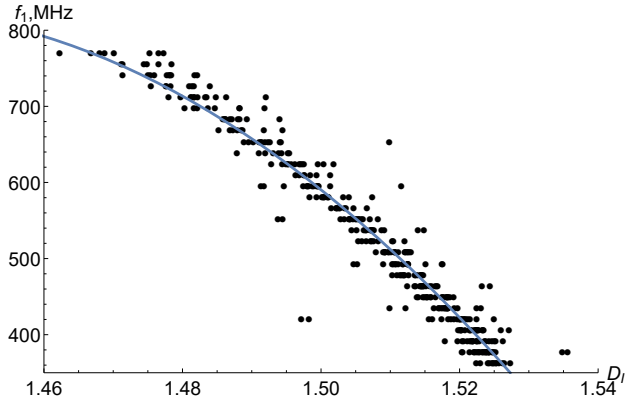


Fig. 5. Ratios of D_I to f_1 for the third iteration

Results of calculations for the correlation dimension (9) are shown in Fig. 6. It can be noted that the points on the correlation field in Figs. 5 and 6 have a very similar allocation.

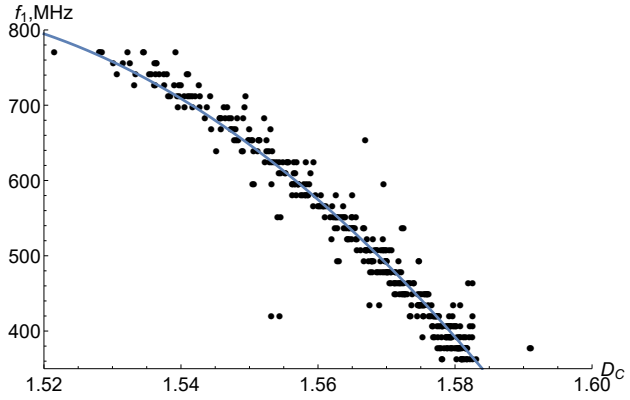


Fig. 6. Ratios of D_C to f_1 for the third iteration

The quadratic regression formula has the following form:

$$\hat{f}_1 = -60297.1\, D_C^2 + 180208\, D_C - 133810. \qquad (14)$$

The formula (14) has the relative error $\delta \approx 3.52\%$. The scatter of points in the Figs 4–6 can be explained by quantization

errors [31], [32] and the choice of the thickness of the line of the curve under study [33].

We also observe that quadratic regression models (12)–(14) can be replaced with linear models. With this replacement, there will be only a slight loss in accuracy.

The Higuchi dimension has the best correlation with f_1. Let us build the regression model for the base frequency:

$$\hat{f}_1 = -1765.194D_H^2 - 697.028D_H + 708.73, \qquad (15)$$

where $\hat{f}_1$ is measured in MHz. Formula (15) approximates the base frequency with the relative error $\delta \approx 1.46\%$. The correlation field for D_H and f_1 and the curve (6) are shown in Fig. 7.

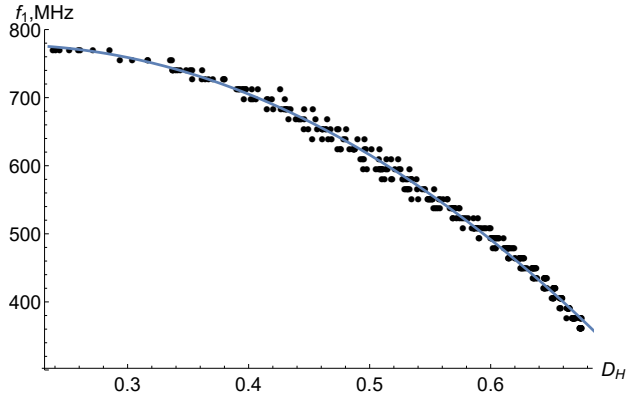


Fig. 7. Ratios of D_H to f_1 for the third iteration

V. Conclusions

Wire Koch-type dipole antennas of the first, second and third iterations are considered. Correlation analysis for the base frequency and several dimensions is carried out. It is established that antennas for all iterations have a strong correlation between the base frequency and the dimensions. Regression formulas with a relative error of 1%–1.5% for the base frequency from the Higuchi dimension D_H for the second and third iterations are obtained.

The obtained regression models can be generalized to the case of antennas with turned arms in space [34] and used in fast algorithms for designing Koch-type wire antennas.

Acknowledgment

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

[1] C. A. Balanis, Antenna theory: analysis and design, 4th ed., John Wiley & Sons, 2016.

[2] C. Singh, V. Grewal, and R. Saxena, "Fractal antennas: a novel miniaturization technique for wireless communications," Int. J. of Recent Trends in Eng., vol. 2, no. 5, 2009, pp. 172–176.

[3] M. H. A. Nasr, "Z-shaped dipole antenna and its fractal iterations," Int. J. Netw. Secur. Appl., vol. 5, 2013, pp. 139–151.

[4] T. A. Milligan, Modern Antenna Design, John Wiley & Sons: Hoboken, 2005.

[5] L. M. Karpukov, V. M. Onufrienko, and S. N. Romanenko, "The properties of the fractal wire antennas," Proc. of MMET Int. Conf., vol. 1, 2002, pp. 310–312.

[6] K. H. Wagh, "A review on fractal antennas for wireless communication," Rev. Electron. Commun. Eng., vol. 32, no. 2, 2015, pp. 37–41.

[7] J. P. Gianvittorio and Y. Rahmat-Samii, "Fractal antennas: A novel antenna miniaturization technique, and applications," IEEE Antennas Propag. Mag., vol. 44, 2002, pp. 20–36.

[8] J. M. Baker and M. F. Iskander, "Electrically small fractal antennas," ISAP, pp. 1242–1243, November 2015 [International Symposium on Antennas and Propagation, 2015].

[9] W. J. Krzysztofik, "Fractal geometry in electromagnetics applications – From antenna to metamaterials," Microw. Rev., vol. 19, 2013, pp. 3–14.

[10] P. Beigi and P. Mohammadi, "A novel small triple-band monopole antenna with crinkle fractal-structure," AEU Int. J. Electron. Commun., vol. 70, 2016, pp. 1382–1387.

[11] J. Anguera, A. Andújar, J. Jayasinghe, V. V. S. S. S. Chakravarthy, P. S. R. Chowdary, J. L. Pijoan, T. Ali, and C. Cattani, "Fractal Antennas: An Historical Perspective," Fractal Fract, vol. 4, no. 3, 2020.

[12] I. Poole and S. Telenius-Lowe, Successful wire antennas, Radio Society of Great Britain, 2011.

[13] C. P. Baliarda, J. Romeu, and A. Cardama, "The Koch monopole: a small fractal antenna," IEEE Trans. Antennas Propag, vol. 48, no. 11, 2000, pp. 1773–1781.

[14] Y. Li, Y. Mi, Y. Wang, and G. Li "The analysis and comparison of the electromagnetic radiation characteristic of the Koch fractal dipole," Proc. of Int. Symp. on Antennas, Propag. & EM Theory, 2012, pp. 15–18.

[15] M. N. A. Karim, M. K. A. Rahim, H. A. Majid, O. Ayop, M. Abu, and F. Zubir, "Log periodic fractal Koch antenna for UHF band applications," Prog. Electromagn. Res, vol. 100, 2010, pp. 201–218.

[16] D. N. Tumakov, G. V. Abgaryan, D. E. Chickrin, and P. A. Kokunin, "Modeling of the Koch-type wire dipole," Appl. Math. Modelling, vol. 51, 2017, pp. 341–360.

[17] P. Banerjee and T.Bezboruah, "Theoretical study of radiation characteristics of short dipole antenna," IMECS, March 2014[International MultiConference of Engineers and Computer Scientists 2014].

[18] G. V. Abgaryan and D. N. Tumakov, "Relation between base frequency of the Koch-type wire dipole, fractal dimensionality and lacunarity," J. of Fund. and Appl. Sciences, vol. -9, no. 1S, 2017, pp. 1885–1898.

[19] G. V. Abgaryan, A. G. Markina, and D. N. Tumakov, "Application of correlation and regression analysis to designing antennas," Revista Publicando, vol. 4, no. 13(2), 2017, pp. -1–13.

[20] G. V. Abgaryan and D. N. Tumakov, "Designing a Koch-type wire antenna by regression analysis," EWDTS, 2018, pp. 504–507 [Proc. of 16th IEEE East-West Design and Test Symp].

[21] A. Markina, D. Tumakov, and N. Pleshchinskii, "Designing the symmetrical eight-tooth-shaped microstrip antenna for Wi-Fi applications", EWDTS, no. 8524698, 2018, pp. 491–495[Proceedings of 2018 IEEE East-West Design and Test Symposium, EWDTS 2018].

[22] D. N. Tumakov, A. G. Markina, and I. B. Badriev, "Fast method for designing a well-matched symmetrical four-tooth-shaped microstrip antenna for Wi-Fi applications," Journal of Physics: Conference Series, vol. 1158, no. 042029.

[23] D. Tumakov, D. Chikrin, and P. Kokunin, "Miniaturization of a Koch-type fractal antenna for Wi-Fi applications," Fractal and Fractional, vol. 4, no. 25, 2020.

[24] Y. Li, Y. Mi, Y. Wang, and G. Li, "The analysis and comparison of the electromagnetic radiation characteristic of the Koch fractal dipole," ISAPE, 2012, pp. 15–18[Proceedings of the International Symposium on Antennas, Propagation and EM Theory 2012].

[25] M. F. Barnsley and A. N. Harrington, "The calculus of fractal interpolation functions," J. Approx. Theory, vol. 57, 1989, pp. 14–34.

[26] K. Igudesman, M. Davletbaev, and G. Shabernev, "New approach to fractal approximation of vector-functions," Abstr. Appl. Anal., vol. 57, 2015, pp. 14–34, no. 278313.

[27] B. Mandelbrot, The fractal geometry of nature, San Francisco: W. H. Freeman, 1982.

[28] F. Moon, Chaotic vibrations: an introduction for applied scientists and engineers, John Wiley & Sons, 2005.

[29] C. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, 1948, pp. 374–423.

[30] T. Higuchi, ”Approach to an irregular time series on the basis of the fractal theory,” Physica D: Nonlinear Phenomena, vol. 31, 1988, pp. 277–283.
[31] K. Foroutan-pour, P. Dutilleul, and D. Smith, ”Advances in the implementation of the box-counting method of fractal dimension estimation,” Applied Mathematics and Computation, vol 105, 1999, pp. 195–210.
[32] D. Da Silva et al, ”A critical appraisal of the box counting method to assess the fractal dimension of tree crowns,” in Advances in Visual Computing, pt. 1, vol. 4291 Lecture Notes in Computer Science, 2nd International Symposium on Visual Computing, eds Bebis G. et al., 2006.
[33] I. Pershin and D. Tumakov, ”On optimal thickness of the curve at calculating the fractal dimension using the box-counting method,” J. of Computational and Theoretical Nanoscience, vol. 16, is. 12, 2019, pp. 5233–5237.
[34] D. Tumakov, A. Ovcharov, D. Chickrin, and P. Kokunin, ”On influence of turning the Koch fractal dipole arms on its base frequency and bandwidth,” Inst. Integr. Omics Appl. Biotechnol, vol. 10, 2019, pp. 28–33.

Miniature Broadband Power Divider in Modern Maritime Communications

Luu Quang Hung
Vietnam Maritime University, Hai Phong, Vietnam.
hung.luuquang@vimaru.edu.vn

***Abstract*—The review of existing miniaturization methods is carried out. The design of a compact broadband power divider is also proposed and its characteristics are calculated. The developed coupler can be used in marine communication systems. The proposed version of the device is technologically simple to manufacture and configure when modeling. At a frequency of 1 GHz, the device has 70% smaller area and 10% narrower frequency band. The review of methods for reducing the device area is performed.**

Keywords—compact, artificial line, quarter-wave length, coupler.

I. INTRODUCTION

Widely used as the basic elements of various microwave – there are passive devices, such as power dividers, designed to divide the signal between output channels in a given ratio. Depending on the purpose of the dividers are equal or unequal division, output to two or more channels. In microwave circuits, it is often necessary to provide power division for branching the path or adding power from multiple sources. As a rule, for these purposes, passive mutual devices are used, which have a fairly simple design, which, due to the principle of reciprocity, can be used both as dividers and adders of microwave power. Dividers-adders are necessary for building balanced mixers and amplifiers, frequency channel separators (multiplexers and demultiplexers), high-power transmitters, multi-element antenna excitation circuits, measuring paths, and so on. Couplers are passive devices designed to split a signal into multiple outputs. Due to their advantages, the couplers are widely used in various radio engineering systems, for example, in communication systems and in measuring radio equipment at sea. The main approach to designing directional couplers is based on the design of individual quarter wave segments, each of which is tuned to the operating frequency. Therefore, at a low operating frequency, the coupler will have large dimensions, which is not always suitable for use in microwave technology. You can find work related to reducing the size of devices operating at low frequencies. Consider some of them [1] - [15], the basic information is shown in Table 1. In [1], bends of lines were used for miniaturization; in [2], compact structures were used. The work [3,6,7,11] describes the process of reducing the area of the device using artificial transmission lines. In [4], symmetric schemes are used as a miniaturization tool, and in [5], a multilayer substrate is used. In [7], the miniaturization of the coupler is based on the coupling between the stubs. In work [9,13] the authors propose to use asymmetric structures allowing to reduce the dimensions of the coupler. The authors in [10] proposed to use high-resistance lines with stubs, which makes it possible to obtain good results on miniaturization. In work [12], periodic capacitive loads are used to obtain a compact coupler. In [14], quasi-lumped elements are used for miniaturization, and in [15], fractals are used for the same purposes. Familiarized with modern theoretical and practical approaches to device design (Table 1). The resolve of this work is to develop and manufacture a microstrip directional coupler that meets the following requirements: Center frequencies 1 GHz and FR4 backing material. The wide use of the power divider is due to the simplicity of its design and implementation.

TABLE I. MINIATURIZATION METHODS EFFICIENCY

	Miniaturization methods	**Relative size, %**	**Center frequency, GHz**
	Conventional microstrip line	100	-
[1]	Bending line	56	3.25
[4]	Symmetric Equivalent Circuits	50	1
[5]	Two layer substrate with rectangular slots in the ground plane	38	2.45
[7]	Source load coupling	30	2.4
[8]	Artificial line segments	25	0.9
[9]	Asymmetric π structures	24	0.9
[10]	High impedance lines and loops	19	0.9
[12]	Periodically capacitive load	49	1.8
[13]	Asymmetrical T-shape structures	45	2.4
[10]	Equivalent Structures	31	1.8
[9]	Electrodynamic structures	30	1.8
[2]	Compact Structure	30.1	1
[3]	Artificial transmission line	28.1	1
[14]	Quasi lumped elements	27.5	0.9
[15]	Fractal technique	24.7	2.4
[6]	Artificial transmission line	21.2	0.9

As you can see from the information in the table, directional couplers when applying various miniaturization methods allows you to get your result. In this paper, the model of the coupler will be described, reduced by replacing quarter wave segments or artificial transmission lines, implemented in print.

II. DESIGN

In the microwave range, distributed elements are typically used to implement directional couplers. The coupler circuit of the microwave power divider in the traditional single layer design is quite widespread in microwave radio systems. When a signal is applied to one of the inputs, the signal passes with equal division into two outputs, and the signal does not arrive at the remaining output. A conventional coupler consists of two parallel quarter wave segments connected by two λ/4 loops, spaced apart by a distance of λ/4. By changing the wave impedances of the lines, it is possible to obtain various power division factors between the outputs of the device. To increase the operating bandwidth of the coupler, it is necessary to extend parallel lines and add loops connecting these lines through λ/4. However, this will be followed by an increase in the total area occupied on the microwave substrate. Figure 1 shows the topology of a broadband coupler implemented on segments with a length of λ/4. The area of such a device with a FR4 substrate is 3748 mm^2. There is a lot of space inside the coupler that is not used in any way and can be effectively used for miniaturization.

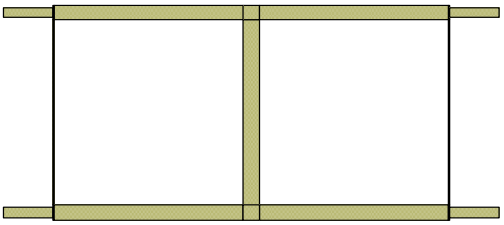


Fig.1. Standard Broadband Coupler Topology

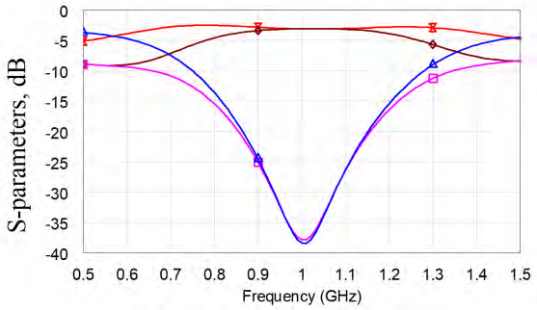


Figure 2. Graph of S-parameters versus standard coupler frequency

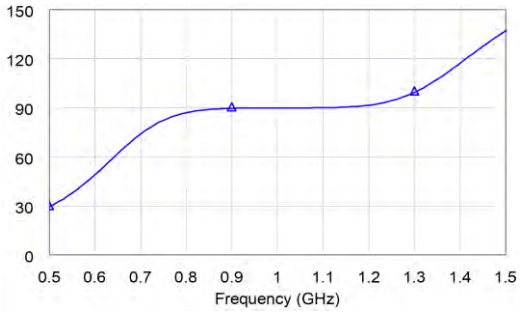


Fig. 3. Graph of the phase

The coupler functions at a central frequency of 1 GHz and has a relative band of 40%. The transmission coefficients are equal at the center frequency and have a value of 3.3 dB. The phase difference corresponds to the theory and is equal to 90 degrees.

To decrease the size of the coupler, it is necessary to calculate the artificial transmission lines. In our case, they will represent the series connected elements of inductance (line with high wave impedance) and capacitance (line with low wave impedance). A comparison of the dimensions and characteristics of conventional segments and artificial lines can be realized in Fig. 4-6. Quarter wave segments included in a conventional directional coupler are easier to bend to a possible length inside the device.

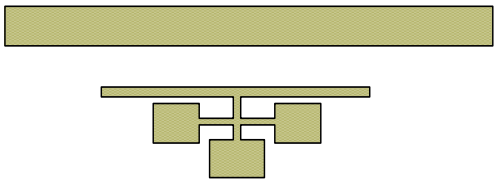


Fig. 4. Comparison of the dimensions of conventional segments and artificial transmission lines

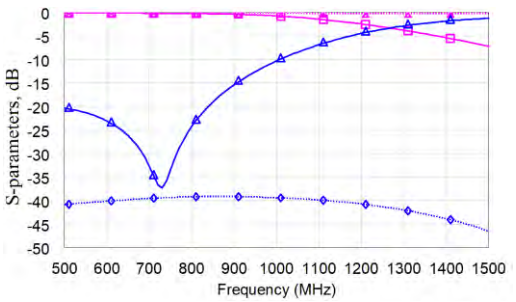


Fig. 5. S-parameters for conventional segments and artificial transmission lines

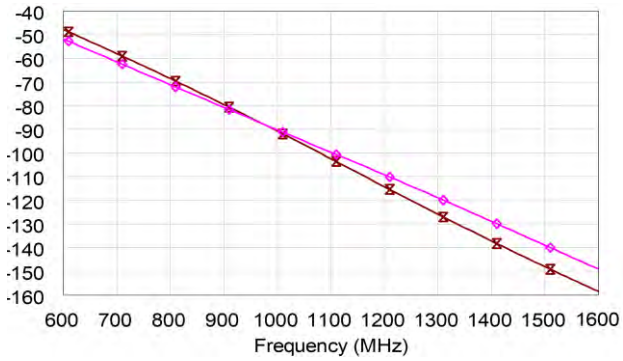


Fig. 6. Phase incursion of conventional segments and artificial transmission lines

Founded on Fig. 4, it can be realized that the artificial transmission lines have a shorter length than ordinary segments, which allows reducing the dimensions of the device. According to Figures 5 and 6, it can be seen that the artificial lines have similar characteristics and that strong changes in the characteristics of the coupler should not occur. Figure 7 shows a compact directional coupler with a wide passband. Its area turned out to be equal to 1172 mm^2, which is 68.7% fewer than the area of a typical coupler. The calculated characteristics of the proposed device are shown in Fig. 8 and 9.

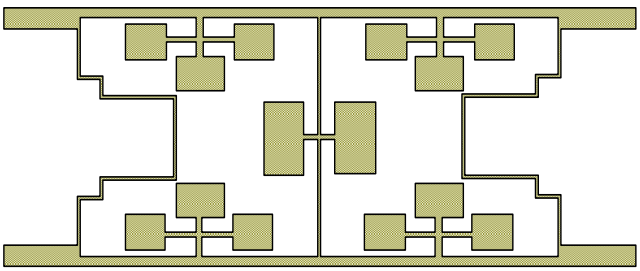


Fig. 7. The topology of a three loop compact coupler

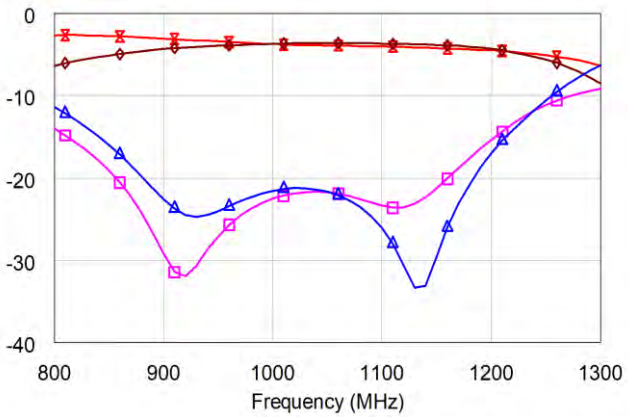


Fig. 8. S-parameter versus frequency for a compact coupler

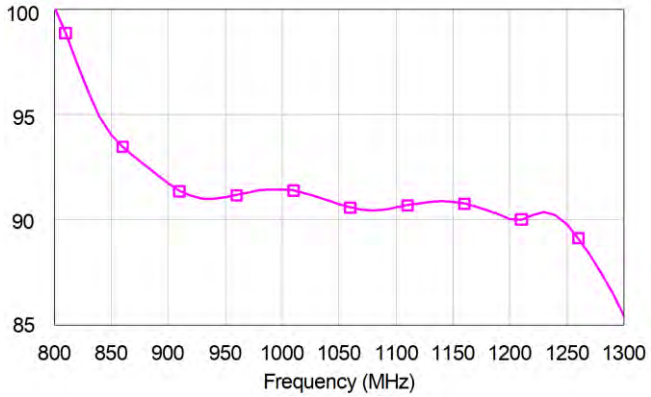


Fig. 9. The phase difference between output of the coupler

The compact coupler functions at a center frequency of 1 GHz and has a relative band of 39%. The transmission coefficients have a slight discrepancy at the center frequency and have values of 3.6 dB and 3.9 dB. The phase difference between the output signals is 90^0. You can see that the compact device operates in a narrower frequency band and the loss in the band has increased. This is all due to the fact that artificial transmission lines are located more densely to each other and spurious phenomena occur. In addition, the coincidence of the characteristics of conventional and artificial lines does not occur in the entire working band of the device. Table 2 shows the data for comparing the couplers.

TABLE II. COMPARATIVE DATA OF RESPONSORS

Parameters	Standard	Compact
bandwidth, MHz	400	390
area, mm^2	3748	1172
Relative area, %	100	31.3
Central frequency, MHz	1000	1000
The phase outputs, ^{o}	90	92

III. CONCLUSION

A compact coupler in marine communication systems having a wide frequency band is proposed in the work. At an operating frequency of 1 GHz, it has an area of 1172 mm^2. As the substrate acts FR4. When replacing conventional lines with artificial ones, it is possible to decrease the scopes of the device, but it is not conceivable to leave the characteristics unchanged. This is due to differences in the characteristics of the structures used in miniaturization from ordinary lines that are part of the traditional coupler. The frequency band was reduced from 400 MHz to 390 MHz. An additional reduction in size is possible, but for this it is necessary to change the geometry of the entire coupler and carry out optimization.

REFERENCES

[1] Ashmi C D., Murmu L. and Dwari S., " A compact branch line coupler using folded microstrip lines," 2013 International Conference on Microwave and Photonics. ICMAP 2013, pp. 1-3, 2013;

[2] Letavin D A., “Miniature microstrip branch line coupler with folded artificial transmission lines,” AEU - International Journal of Electronics and Communications, vol. 99, pp. 8-13, 2019;

[3] Letavin D A., “Two Methods for Miniaturization of Stub Quadrature Couplers,” Journal of Communications Technology and Electronics, vol. 63, Issue 8, pp. 933-935, 2018;

[4] Ahn H R. and Nam S., «Compact Microstrip 3-dB Coupled-Line Ring and Branch-Line Hybrids With New Symmetric Equivalent Circuits,» IEEE Transactions on Microwave Theory and Techniques, vol. 61, Issue 3, 1067–1078;

[5] Ausordin S F., Rahim S K A., Seman N. and Dewan R., “A compact 3-dB coupler on a dual substrate layer with a rectangular slotted microstrip ground plane,” 2013 IEEE Business Engineering and Industrial Applications Colloquium, pp. 156-160, BEIAC 2013;

[6] Letavin D A., “Microstrip directional coupler with reduced dimensions due to the use of compact structures,” 2019 International Multi-Conference on Engineering, Computer and Information Sciences, pp. 43-46, SIBIRCON 2019;

[7] Lin T W., Wu J Y., and Kuo JT., “Compact filtering branch-line coupler with source-load coupling,” 2016 IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition, IWEM 2016.

[8] Ostankov AV., “Microstrip directional coupler made on the basis of segments of artificial long lines,” Moscow: Scientific Technologies. Series "Natural and Technical Sciences", № 1. -2016;

[9] Letavin D A. and Abdullin R R., "Development of Compact Coupler Devices on Electrody-Namic Structures with Different Thickness Substrates," 2018 IEEE Radio and Antenna Days of the Indian Ocean (RADIO), Grand Port, 2018, pp. 1-2, doi: 10.23919/RADIO.2018.8572387;

[10] Letavin D. A. and Shabunin S. N., "Investigation of the construction of the phase shifter based on a 3-dB directional coupler," 2017 Panhellenic Conference on Electronics and Telecommunicatio

[11] Phani K., Barik R K. and Karthikeyan S S., "A novel two section branch line coupler employing different transmission line techniques". AEU - International Journal of Electronics and Communications, vol. 70, Issue 5, 738–742;

[12] Qiuyi W., Yimin Y., Ying W., Xiaowei S., and Ming Y., “General model for loaded stub branch line coupler,” 2016 IEEE MTT-S International Microwave Symposium, IMS 2016;

[13] Koziel S. and Bekasiewicz A., "Novel structure and size-reduction-oriented design of microstrip compact rat-race coupler," 2016 IEEE/ACES International Conference on Wireless Information Technology and Systems (ICWITS) and Applied Computational Electromagnetics (ACES) 2016;

[14] Eccleston K W. and Ong S H., “Compact planar microstripline branch-line and rat-race coupler,” IEEE Trans. Microw. Theory Tech., vol. 51, pp. 2119–2125;

[15] Liao S S., Sun P T., Chin N C. and Peng J T., “A novel compact-size branch-line coupler ,” IEEE Microw. Wireless Compon. Lett., vol. 15, pp. 588–590;

;

Solving Problem of Electromagnetic Wave Diffraction by a Metal Plate Using CUDA

Dinara Giniyatova
Institute of Comp. Math. and Information Technologies
Kazan Federal University
Kazan, Russia
normaliti@gmail.com

Dmitrii Tumakov
Institute of Comp. Math. and Information Technologies
Kazan Federal University
Kazan, Russia
dtumakov@kpfu.ru

Angelina Markina
Institute of Comp. Math. and Information Technologies
Kazan Federal University
Kazan, Russia
m8angelin@gmail.com

***Abstract*—In the present paper the problem of plane electromagnetic wave diffraction by a thin metal plate is considered. A numerical algorithm is developed using method of moments with NVIDIA CUDA technology implementation. The results of numerical modeling of a plane wave diffraction by the square thin metallic plate is shown. Comparative analysis of the performance for CPU and GPU is carried out. It is shown that the method of moments implementation by graphical processor provides a sufficient gain in the performance.**

***Index Terms*—electromagnetic wave, diffraction, integral equation, method of moments, CUDA technology**

I. Introduction

The problems of diffraction of electromagnetic waves arise in the study of various kinds of complex electrodynamic systems. For their analysis, it is necessary to use strict analytical methods of applied electrodynamics [1]–[3] or approximate numerical methods [4]–[6]. To date, the following methods are widely used in specialized software: the moment method (MoM), the finite element method (FEM) [7], and the finite difference method in the time domain (FDTD) [8]. All these methods lead to the need to solve complex systems of linear algebraic equations, the order of which directly depends on the desired degree of accuracy of solving the problem. The use of effective numerical methods and new computer technologies make it possible to solve similar problems within an acceptable time. A promising technology, from the point of view of calculation time, is parallel computing on a graphics processor (NVIDIA CUDA) [9]–[12]. In the present paper, we consider a parallel algorithm for solving the diffraction problem by the method of moments on CUDA.

As is known [13], [14], the problem of diffraction of electromagnetic waves on a perfectly conducting surface can be described by the operator equation for the surface current. The operator can be a linear integral or integro-differential operator, and integration is carried out over the entire diffraction surface. To solve such equations, the method of moments is used. Harrington and his monograph "Field Computation by Moment Methods" described the method of moments most fully [15]; the current state of the method of moments in electrodynamics problems is described in the monographs by Sadiku [16] and Gibson [14]. In addition, a number of works are devoted to the mathematical justification of this method and the convergence of the approximate solution to the exact one [17]. As already noted above, the separation of the diffraction surface into small finite regions leads to the construction and further numerical solution of systems of linear algebraic equations of a very high order. On the other hand, this class of tasks lends itself well to parallelization, and the architecture of the graphic processor (GPU) is well optimized for parallel data processing.

II. Diffraction Problem Statement

We consider the problem of electromagnetic field diffraction on an perfectly conducting thin plate of an arbitrary shape (see, for example, [18]). Let $\Omega \subset R^2$ be a bounded domain with a piecewise-smooth boundary Γ consisting of a finite number of arcs of the class C^∞ converging at non-zero angles. The problem of diffraction of an external electromagnetic field $\mathbf{E}^0, \mathbf{H}^0$ on a perfectly conducting plate Ω, located in free space with a wave number $k, k^2 = \omega^2\epsilon\mu$, consists in the determining scattered electromagnetic field

$$\mathbf{E}, \mathbf{H} \in C^2(R^3 \setminus \overline{\Omega}) \bigcap_{\delta>0} C(\overline{R}^3_+ \setminus \Gamma_\delta) \bigcap_{\delta>0} C(\overline{R}^3_- \setminus \Gamma_\delta) \quad (1)$$

satisfying homogeneous Maxwell equations:

$$\begin{aligned} \operatorname{Rot}\mathbf{H} &= -ik\mathbf{E}, \\ \operatorname{Rot}\mathbf{E} &= ik\mathbf{H}, \quad \mathbf{x} \in R^3 \setminus \overline{\Omega} \end{aligned} \quad (2)$$

boundary conditions for tangent components of the electric field on the plate surface:

$$\mathbf{E}_\tau|_\Omega = -\mathbf{E}^0_\tau|_\Omega \quad (3)$$

conditions of finite energy in any limited amount of space:

$$\mathbf{E}, \mathbf{H} \in L^2_{loc}(R^3) \quad (4)$$

and conditions at infinity:

$$\begin{aligned} \frac{\partial}{\partial r}\begin{pmatrix}\mathbf{E}\\ \mathbf{H}\end{pmatrix} - ik\begin{pmatrix}\mathbf{E}\\ \mathbf{H}\end{pmatrix} &= o(r^{-1}), \\ \begin{pmatrix}\mathbf{E}\\ \mathbf{H}\end{pmatrix} &= O(r^{-1}), \quad r : |\mathbf{x}| \to \infty. \end{aligned} \quad (5)$$

For the full field, $\mathbf{E}^{tot} = \mathbf{E}^0 + \mathbf{E}, \quad \mathbf{H}^{tot} = \mathbf{H}^0 + \mathbf{E}$. We assume that all sources of the incident field are outside of the plate $\overline{\Omega}$ so that for some $\delta > 0$

$$\mathbf{E}^0 \in C^\infty(\Omega_\delta), \quad \Omega_\delta = \{\mathbf{x} : |\mathbf{x} - \mathbf{y}| < \delta, \mathbf{y} \in \Omega\} \tag{6}$$

whence it follows that

$$\mathbf{E}^0_\tau|_\Omega \in C^\infty(\overline{\Omega}). \tag{7}$$

Often, either a plane wave or an electric or magnetic dipole located outside of $\overline{\Omega}$ is considered as an incident field. In this case, conditions (6), (7) are satisfied. The field $\mathbf{E}^0, \mathbf{H}^0$ is a solution to the system of Maxwell equations in free space without a plate.

One approach to solving problem (1)-(7) is to reduce it to an integro-differential equation on a plate [13]. This method is often called the surface current method.

Now let S be the open surface of a perfectly conducting plate with the unit normal $\mathbf{n}$. By $\mathbf{E}^i$ we denote the electric field defined to be the field due to a source in the absence of a plate. It induces a surface currents $\mathbf{J}$ on S. Since S is an open surface, we consider $\mathbf{J}$ as the sum of the surface currents on opposite sides of S and, therefore, the normal component $\mathbf{J}$ should vanish on boundaries of S. The scattered electric field $\mathbf{E}^s$ can be computed by the formula [14]

$$\mathbf{E}^s = -i\omega\mathbf{A} - \nabla\Phi, \tag{8}$$

where $\mathbf{A}$ and Φ are the vector and scalar potentials, respectively. It is known [14] that the potentials are related to the exciting current through the Green's function. In free space, the following formulas are valid

$$\mathbf{A}(\mathbf{r}) = \mu \int\limits_S \mathbf{J}(\mathbf{r}')G(\mathbf{r}, \mathbf{r}')dS', \tag{9}$$

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon} \int\limits_S \sigma G(\mathbf{r}, \mathbf{r}')dS', \tag{10}$$

where Green's function defined as

$$G(\mathbf{r}, \mathbf{r}') = \frac{e^{-ik|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r} - \mathbf{r}'|},$$

$k = \omega\sqrt{\mu\varepsilon} = 2\pi/\lambda$ (λ is a wavelength) and $|\mathbf{r} - \mathbf{r}'|$ is the distance between the arbitrarily located observation point $\mathbf{r}$ and the source point $\mathbf{r}'$ on S. The surface charge density σ is related to the surface divergence of current through the equation of continuity

$$\nabla_s \cdot \mathbf{J} = -i\omega\sigma. \tag{11}$$

The boundary condition for the electric field in the case of the perfectly conducting surface is

$$\mathbf{n} \times (\mathbf{E}^i + \mathbf{E}^s) = 0, \tag{12}$$

whence, using (8) we obtain the integro-differential equation with respect to $\mathbf{J}$

$$-\mathbf{E}^i_{tan} = (-i\omega\mathbf{A} - \nabla\Phi)_{tan}, \quad \mathbf{r} \in S. \tag{13}$$

Together with (9)-(11), equation (13) is the so-called the electric field integral equation (EFIE). In the literature, the classic EFIE formulation for perfectly conductive surfaces is written as

$$\mathbf{n} \times \int_S G(\mathbf{r}, \mathbf{r}') \left[\mathbf{J}(\mathbf{r}') + \frac{1}{k^2}\nabla(\nabla \cdot \mathbf{J}(\mathbf{r}'))\right] dS' = \frac{1}{ik\eta_0}\mathbf{n} \times \mathbf{E}^i(\mathbf{r}),$$

and equation (13) is called the equation in terms of mixed potentials (Mixed Potential Integral Equation). Nevertheless, hereinafter, we will use the term EFIE, implying equation (13), taking into account (9)-(11).

III. The Method of Moments

The method of moments (see [14], [15]) is one of the most common numerical methods of electrodynamics, used to calculate surface currents on plane metal or dielectric structures when emitted in free space. It is used for analysis and modeling of flat structures that allow the inclusion of dielectrics. In fact, the method of moments is a way to solve Maxwell's equations written in integral form (EFIE, MFIE) in the frequency domain. The main advantage of the method is that only the metal elements of the object under consideration are discretized, and the distribution of the surface current on the metal acts as an unknown quantity. This distinguishes the method of moments from other methods of EM modeling, where in addition to the object itself, some limited space around is discretized, as, for example, when solving problems of finding a field in a volume. We describe the main stages of the method of moments. As mentioned earlier, the method of moments is a method of solving an operator problem in the form

$$Lf = g, \tag{14}$$

where L is a linear operator, g is a known perturbation function, f is an unknown function. In our case, L is an integro-differential operator, f is an unknown current function $\mathbf{J}$, and g is a known excitation source (incident field $\mathbf{E}^i$) . We represent the function f as a finite sum of N basis functions f_n with unknown weight coefficients α_n:

$$f \approx \sum_{n=1}^{N} \alpha_n f_n. \tag{15}$$

Then, due to the linearity of the operator L,

$$\sum_{n=1}^{N} \alpha_n L(f_n) \approx g, \tag{16}$$

and the residual R is defined as

$$R = g - \sum_{n=1}^{N} \alpha_n L(f_n). \tag{17}$$

The basis functions f_n are chosen in such a way as to correctly model the expected properties of the unknown function f.

Next, we define the scalar product, or *moment* between the basis functions $f_n(r')$ and the test functions $g_m(r)$ as

$$\langle g_m, f_n \rangle = \int_{g_m} g_m(\mathbf{r}) \cdot \int_{f_n} f_n(\mathbf{r}'), \quad m = \overline{1, N}, \qquad (18)$$

where the presented integrals can be linear, surface, or volumetric depending on the type of basis and test functions. We require that the scalar product of each test function with the residual function be zero, then

$$\sum_{n=1}^{N} \alpha_n \langle g_m, L(f_n) \rangle = \langle g_m, g \rangle, \quad m = \overline{1, N}. \qquad (19)$$

The equality (19) is a system of linear algebraic equations for unknown coefficients α_n and in matrix form can be written as $\mathbf{Z}\,\mathbf{a} = \mathbf{b}$, where

$$\mathbf{Z} = \begin{pmatrix} \langle g_1, L(f_1) \rangle & \langle g_1, L(f_2) \rangle & ... & \langle g_1, L(f_N) \rangle \\ \langle g_2, L(f_1) \rangle & \langle g_2, L(f_2) \rangle & ... & \langle g_2, L(f_N) \rangle \\ ... & ... & & \\ \langle g_N, L(f_1) \rangle & \langle g_N, L(f_2) \rangle & ... & \langle g_N, L(f_N) \rangle \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} \langle g_1, g \rangle \\ \langle g_2, g \rangle \\ ... \\ \langle g_N, g \rangle \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_N \end{pmatrix}.$$

Note that in the method of moments the resulting matrix of the system is completely filled, in contrast, for example, to the finite element method, where the matrix of the system is, as a rule, very sparse and symmetrical. Solving (19), we determine the unknown coefficients α_n by which the desired function f is reconstructed. Thus, $f = \langle \bar{\mathbf{f}}, \mathbf{Z}^{-1}\mathbf{b} \rangle$, $\bar{\mathbf{f}} = (f_1, f_2, ..., f_N)^T$. This completes the procedure of the method of moments.

IV. Basis and Test Functions

One of the most popular basis functions used in calculating the surface current are the so-called RWG functions proposed in [19]. They are conveniently used to search for an approximate EFIE solution when the surface of a perfectly conducting body is divided into elementary triangular patches. We will use standard terms, such as a face, to denote the surface of an elementary triangular patch, an edge (boundary edge) to indicate one of its sides, and a vertex to indicate the vertices of a triangle.

First of all, we note that each basis RWG function is associated with one inner edge and vanishes everywhere on S, except for a pair of triangles adjacent to this edge. Fig. 1 shows two such triangles, T_n^+ and T_n^-, adjacent to the n-th edge. Points belonging to the triangle T_n^+ can be described both in global coordinates by the radius vector $\mathbf{r}$, and in local coordinates using the radius vector ρ_n^+ defined relative to the free vertex of the triangle T_n^+. A similar remark is also true for the triangle T_n^- with the only difference being that the vector ρ_n^- is directed from the point belonging to the triangle to the free vertex T_n^-. The choice of "positive" and "negative" triangles is arbitrary, given that for the entire cycle of calculating the surface current, it will not change.

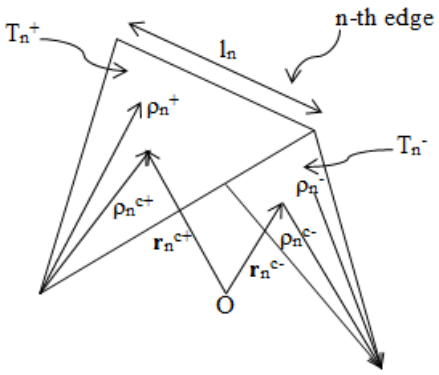


Fig. 1. Triangle pair and RWG parameters associated with inner edge n.

Basis function associated with the n-th inner edge defined as:

$$f_n(r) = \begin{cases} \dfrac{l_n}{2A_n^+}\rho_n^+, & r \in T_n^+, \\ \dfrac{l_n}{2A_n^-}\rho_n^-, & r \in T_n^-, \\ 0, & \text{otherwise}, \end{cases} \qquad (20)$$

where l_n is the length of the n-th edge, A_n^+ and A_n^- are the areas of the triangles T_n^+ and T_n^-, respectively. The properties of RWG functions are described in detail in [19]. Following the method of moments, we represent the surface current everywhere on S in the form of an approximate formula

$$\mathbf{J} \approx \sum_{n=1}^{N} \alpha_n f_n(\mathbf{r}), \qquad (21)$$

where N is the number of inner edges.

The next step in the method of moments is the testing procedure or multiplying the original equation by testing functions. Generally speaking, for the testing procedure, it is permissible to use any functions. However, their choice for a specific problem is crucial. One of the most effective methods is *Galerkin's method*, when the same basis functions are chosen as test functions. This ensures that the boundary conditions are observed throughout the solution area, and not just at discrete points. Therefore, we take the same RWG functions as test functions. We define scalar product as $\langle f, g \rangle = \int_S f \cdot g\, dS$ and test the equation (13) by the RWG functions. We obtain

$$\langle \mathbf{E}^i, f_m \rangle = i\omega \langle \mathbf{A}, f_m \rangle + \langle \nabla\Phi, f_m \rangle. \qquad (22)$$

Using methods for calculating the surface integral and the f_m property at the S boundaries, the last term in (22) can be written as

$$\langle \nabla\Phi, f_m \rangle = -\int_S \Phi\, \nabla_s \cdot f_m dS. \qquad (23)$$

Then, using

$$\nabla_s \cdot f_n = \begin{cases} \dfrac{l_n}{A_n^+}, & \mathbf{r} \in T_n^+, \\ -\dfrac{l_n}{A_n^-}, & \mathbf{r} \in T_n^-, \\ 0, & \text{otherwise}, \end{cases}$$

the integral in (23) can be approximated as follows

$$\int_S \Phi \nabla_s \cdot f_m dS = l_m \left(\frac{1}{A_m^+} \int_{T_m^+} \Phi dS - \frac{1}{A_m^-} \int_{T_m^-} \Phi dS \right) \cong l_m [\Phi(\mathbf{r}_m^{c+}) - \Phi(\mathbf{r}_m^{c-})]. \quad (24)$$

In (24), the average value of Φ for each triangle was replaced by the value Φ in the center of mass of the triangles. Using similar arguments, we can approximate the terms in (22) containing the vector potential and the incident field. We show this by the example of the term $\langle \mathbf{E}^i, f_m \rangle$:

$$\langle \mathbf{E}^i, f_m \rangle = \int_S \mathbf{E}^i \cdot f_m dS = \frac{l_m}{2} \left(\frac{1}{A_m^+} \int_{T_m^+} \mathbf{E}^i \cdot \rho_m^+ dS + \frac{1}{A_m^-} \int_{T_m^-} \mathbf{E}^i \cdot \rho_m^- dS \right) \cong \frac{l_m}{2} \left(\mathbf{E}^i(\mathbf{r}_m^{c+}) \rho_m^{c+} + \mathbf{E}^i(\mathbf{r}_m^{c-}) \rho_m^{c-} \right). \quad (25)$$

Thus, applying the testing procedure for EFIE by RWG functions, with (23)-(25), we obtain an equation

$$i\omega l_m \left[\mathbf{A}(\mathbf{r}_m^{c+}) \frac{\rho_m^{c+}}{2} + \mathbf{A}(\mathbf{r}_m^{c-}) \frac{\rho_m^{c-}}{2} \right] + l_m [\Phi(\mathbf{r}_m^{c+}) - \Phi(\mathbf{r}_m^{c-})] = l_m \left[\mathbf{E}^i(\mathbf{r}_m^{c+}) \frac{\rho_m^{c+}}{2} + \mathbf{E}^i(\mathbf{r}_m^{c-}) \frac{\rho_m^{c-}}{2} \right] \quad (26)$$

that is valid for each inner edge, $m = \overline{1, N}$.

V. Filling the Moment Matrix and SLAE

After inserting the expansion for the surface current (21) into equation (26), we obtain a system of linear algebraic equations (SLAE) of size $N \times N$, which can be represented as

$$\mathbf{Z}\,\mathbf{I} = \mathbf{V}, \quad (27)$$

where $\mathbf{Z} = [Z_{mn}]$ is the $N \times N$ matrix, $\mathbf{I} = [\alpha_n]$ is the column of unknown coefficients, $\mathbf{V} = [V_m]$ is the column of the known right-hand side. The elements of the matrix $\mathbf{Z}$ and the column $\mathbf{V}$ are determined by the following formulas:

$$Z_{mn} = l_m \left[i\omega \left(A_{mn}^+ \cdot \frac{\rho_m^{c+}}{2} + A_{mn}^- \cdot \frac{\rho_m^{c-}}{2} \right) + \Phi_{mn}^- - \Phi_{mn}^+ \right], \quad (28)$$

$$V_m = l_m \left(\mathbf{E}_m^+ \cdot \frac{\rho_m^{c+}}{2} + \mathbf{E}_m^- \cdot \frac{\rho_m^{c-}}{2} \right), \quad (29)$$

where

$$A_{mn}^{\pm} = \frac{\mu}{4\pi} \int_S f_n(\mathbf{r}') \frac{e^{-ik|\mathbf{r}_m^{c\pm} - \mathbf{r}'|}}{|\mathbf{r}_m^{c\pm} - \mathbf{r}'|} dS', \quad (30)$$

$$\Phi_{mn}^{\pm} = -\frac{1}{4\pi\epsilon i\omega} \int_S \nabla_s' f_n(\mathbf{r}') \frac{e^{-ik|\mathbf{r}_m^{c\pm} - \mathbf{r}'|}}{|\mathbf{r}_m^{c\pm} - \mathbf{r}'|} dS', \quad (31)$$

$$E_m^{\pm} = \mathbf{E}^i(\mathbf{r}_m^{c\pm}). \quad (32)$$

After defining the elements of the moment matrix $\mathbf{Z}$ and the vector $\mathbf{V}$, we can solve the resulting system (27) with respect to the vector of unknown coefficients α_n by one of the well-known methods for solving SLAEs.

VI. Diffraction Problem Solution on GPU

The numerical solution of the problem can be conditionally divided into three main stages. At the first stage, we build triangular grid of the plate surface and the array of RWG elements. At the second stage, we compute the elements of the moment matrix and derive the final SLAE; at the third stage, we solve the SLAE and build the required function. The most time-consuming stage is the filling of the moment matrix. However, as can be seen from formulas (28)-(31), each element of the matrix can be calculated independently. Thus, the task is easy to parallelize. When launching the main computing core, which is responsible for calculating the elements of the moment matrix, a two-dimensional grid of blocks and two-dimensional blocks were used, since the matrix of the final SLAE is presented in memory as a two-dimensional data array.

The calculation program is written in the programming language C++ and CUDA C, provided by NVIDIA, which implements support for the CUDA API for compiling code that runs on a GPU. For calculations, we used a personal computer with the Intel Core i3-5005U processor (2 GHz), RAM is 4 GB with the graphics accelerator GeForce 920M. Fig. 2 shows the distribution graphs of the normal component of the current $|\mathbf{J}_x|$ on a square plate along sections parallel to the Ox and Oy axes with sides equal to $0.15 * \lambda$ and λ, respectively, and with different level of discretization of the area: 98 and 450 triangles, respectively, which corresponds to 133 and 645 RWG elements, respectively.

The obtained graphs fully correspond to the graphs presented in [18], [19]. The performance gain when filling out the matrix of moments with respect to calculations on a single CPU core is presented in Table I. The implemented program allows to read any rectangular plates; a square plate was chosen to compare the results.

Our results are in good agreement with the earlier results obtained by other authors. Fig. 3 shows the dependence of acceleration on the "size" of the original problem, i.e. on the number of RWG elements. It can be seen that the implementation of this stage of the method of moments on the GPU gives

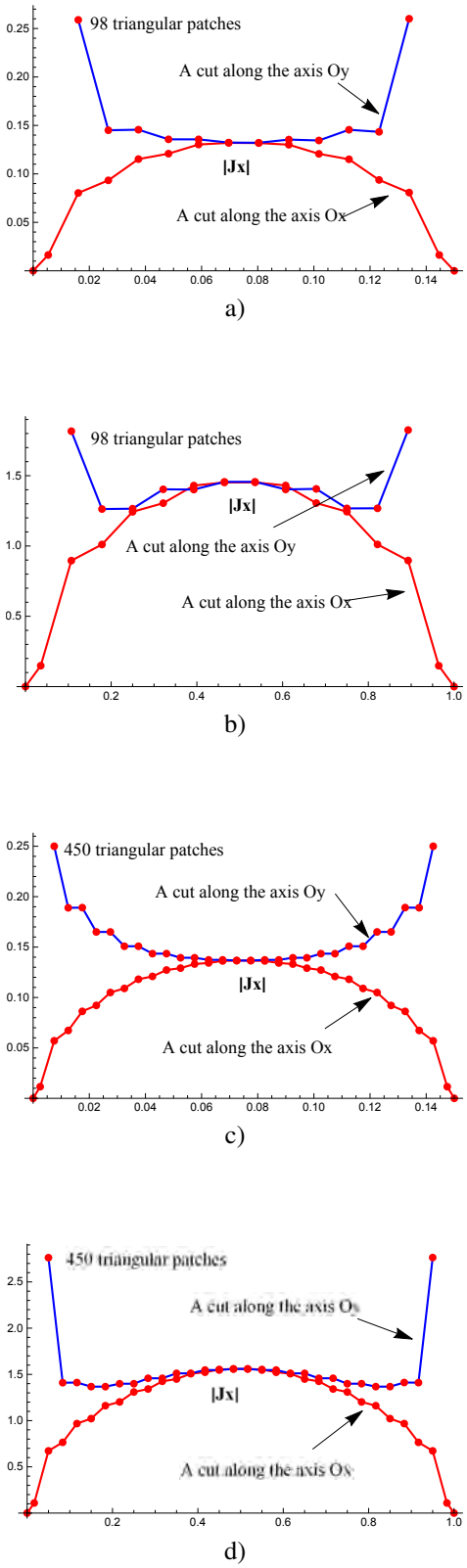


Fig. 2. Distribution of normal component of current on a) 0.15λ square plate, number of triangular patches is 98, b) λ square plate, number of triangular patches is 98, c) 0.15λ, square plate, number of triangular patches is 450, d) λ square plate, number of triangular patches is 450.

TABLE I
TIME REQUIRED FOR COMPUTING ON CPU AND GPU.

Number of RWG elements	Matrix Fill Time (s)		Acceleration
	CPU	GPU	
133	0.88	0.61	1.42
280	3.80	0.79	4.81
560	15.19	1.40	10.84
1045	52.83	3.43	15.40
1408	95.63	5.95	16.06
1825	159.84	9.92	16.08

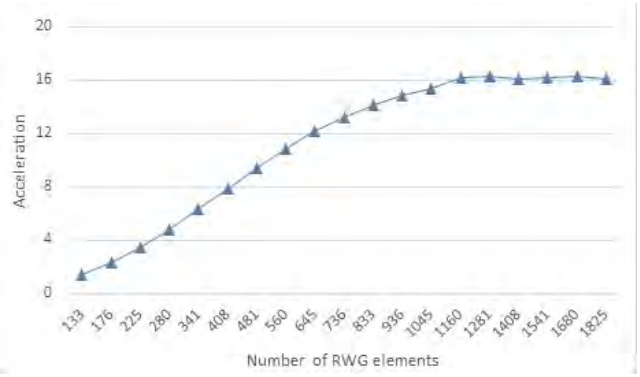


Fig. 3. Dependence of acceleration on the number of RWG elements

higher performance indicators than the implementation of the same algorithm on the CPU even taking into account the need to copy data from the CPU to the GPU and vice versa. The maximum acceleration reaches ≈ 16.

VII. CONCLUSIONS

A parallel algorithm is implemented on CUDA to solve the diffraction problem on a metal perfectly conducting plate. Numerical calculations are performed for the case of a rectangular plate. Conclusions are made about the good agreement of the results of calculations with previous works.

An almost 16-fold the acceleration at the stage of forming the moment matrix (with copying data from the CPU to the GPU and vice versa) on a not high-performance video card is obtained. This value can be increased due to the absence of copying the moment matrix from device to host in solving a SLAE on a GPU. Thus, calculations on less-powerful GPUs built in laptops seem to be an effective tool for hardware acceleration of the method of moments in solving various electrodynamic problems.

ACKNOWLEDGMENT

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

[1] T. Nagasaka and K. Kobayashi, "Wiener-Hopf analysis of the plane wave diffraction by a thin material strip," IEICE Transactions on Electronics, vol. E100-C, no. 1, 2017, pp. 11-19.

[2] E. Luz, E. Granot, and B. A. Malomed, "Analytical boundary-based method for diffraction calculations," Journal of Optics, vol. 22, 2019, 025601.
[3] M. A. Nethercote, R. C. Assier, and I. D. Abrahams, "Analytical methods for perfect wedge diffraction: A review," Wave Motion, vol. 93, 2020, 102479.
[4] D. N. Tumakov and A. R. Tukhvatova, "Diffraction of an electromagnetic wave by gaps between plates," Lobachevskii Journal of Mathematics, vol. 33, no. 4, 2012, pp. 392-401.
[5] D. N. Tumakov, "Iterative method for solving the problem of scattering of an electromagnetic wave by a partially shielded conducting sphere," Applied Mathematical Sciences, vol. 8, no. 118, 2014, pp. 5887-5898.
[6] I. T. Selezov, Y. G. Kryvonos, I. S. Gandzha, "Some analytical and numerical methods in the theory of wave propagation and diffraction," In: Wave Propagation and Diffraction. Foundations of Engineering Mechanics. Springer, Singapore, 2018, pp. 1-24.
[7] K. Shibata and M. Kobayashi, "Difference between the method of moments and the finite element method for estimation of complex permittivity in liquids using a coaxial probe," Int. Symp. Electromagnetic Comp., Barcelona, Spain, 2019, pp. 100-105.
[8] A. Taflove and S. Hagness, Computational Electrodynamics: the Finite-Difference Time-Domain Method, 2nd ed., Artech House, 2000.
[9] I. P. Molostov and V. V. Scherbinin, "Application of NVIDIA CUDA technology for numerical simulation of electromagnetic pulses propagation," Izvestiya of Altai State University, vol. 85, no. 1, 2015, pp. 39-43.
[10] Y. Zhang, X. Mei, and H. Lin, "OpenMP-CUDA accelerated moment method for homogeneous dielectric objects," IEEE Antennas and Propagation Society, AP-S International Symposium (Digest), 2014, pp. 1634-1635.
[11] J. Li, Z. Zhang, K. Zheng, C. Wu, and S. Gao, "GPU accelerated non-illuminated graphical electromagnetic computing method with high accuracy," Optik-International J. Light and Electron Optics, vol. 142, 2017, pp. 523-528.
[12] S. Peng and Z. P. Nie, "Acceleration of the method of moments calculations by using graphics processing units," Antennas and Propagation, IEEE Transactions, vol. 56, no. 7, 2008, pp. 2130-2133.
[13] C. A. Balanis, Antenna Theory: Analysis and Design, 4th ed., John Wiley & Sons, 2016.
[14] W. C. Gibson, The Method of Moments in Electromagnetics, Taylor & Francis Group, 2008.
[15] R. F. Harrington, "Field computation by moment methods," IEEE Trans. Antennas Propagat., vol. 4, no. 9, 1993, pp. 229-231.
[16] M. N. O. Sadiku, Elements of Electromagnetics, New York: Oxford University Press, 2001.
[17] R. Mittra and C. A. Klein, Stability and convergence of moment method solutions. In: Mittra R. (eds) Numerical and Asymptotic Techniques in Electromagnetics, Topics in Applied Physics, vol. 3. Springer, Berlin, Heidelberg, 1975.
[18] Y. G. Smirnov, M. Y. Medvedik, and M. A. Maksimova, "The solution of the problem of electromagnetic wave diffraction on screens of complex shape," Izvestiya Vysshikh Uchebnykh Zavedenii. Volga region. Physics and Mathematics, no. 4, 2012, pp. 59-72.
[19] M. S. Rao, R. D. Wilton, and A .W. Glisson, "Electromagnetic scattering by surfaces of arbitrary shape," IEEE Trans. Antennas Propagation, vol. AP-30, no. 3, 1982, pp. 409-418.

Convolution Neural Network Learning Features for Handwritten Digit Recognition

1st Zufar Kayumov
Institute of Computational Mathematics and Information Technologies
Kazan Federal University
Kazan, Russia
kayumov.zufar@gmail.com

2nd Dmitrii Tumakov
Institute of Computational Mathematics and Information Technologies
Kazan Federal University
Kazan, Russia
dtumakov@kpfu.ru

***Abstract*— The classical six-layer neural network is considered. This network is used to recognize handwritten digit patterns from the MNIST database. The influence of the size of mini-batches on the learning speed and pattern recognition accuracy is analyzed. The optimal sizes of mini-batch are obtained. The relationship between the accuracy of training and the accuracy of test samples is considered. The change in the values of the radius vector of the scales during training is shown. Conclusions are drawn about the influence of the initial value of the balance on the recognition accuracy. A more accurate formula is obtained for the limits, in which the initial values of the weights of the neural network are generated.**

Keywords—convolutional neural network, handwritten digits, recognition, neural network training.

I. INTRODUCTION

Currently, neural networks play one of the main roles in our life. Neural networks are widely used in science [1, 2], technology [3, 4] and many other areas [5, 6]. One of the applications of neural networks is image processing. This includes various classification, localization and pattern recognition tasks. The most common class among networks for these tasks is the class of convolutional neural networks [7-11]. Here, thanks to the use of convolutional layers, the input data is filtered from unnecessary details. This allows further processing only useful information, due to which there is an effective recognition of objects. Objects in pattern recognition are understood as various objects and phenomena, processes and situations, signals, etc. For example, convolutional neural networks are used to recognize objects and symbols [12]. Also for recognition of human actions [13], facial emotions [14, 15], pedestrian detection [16].

The convolutional neural network is a class of deep neural networks that is often used in image analysis. The main idea of convolutional neural networks is to use alternating convolutional and subsampling layers and a multilayer perceptron at the output.

Keras was chosen as the framework for writing a convolutional neural network program in Python.

In present work, we will consider the recognition of handwritten digits from the MNIST database [17] by a convolutional neural network. Previously, various authors have repeatedly made numerous attempts to achieve maximum accuracy of handwritten digit recognition. For example, when using a single-level perceptron, the error was 12%, and for two-layer networks using elastic deformations, the error reached 0.7% [18]. However, the best results were obtained only when using deep convolutional neural networks. For example, a set of 35 six-layer convolutional neural networks with preprocessing and wide normalization for training showed 0.23% error [19], while using an ensemble of five such networks with a significantly expanded dataset, the error rate was only 0.21% [20].

However, all of these techniques do distortion or image processing. Such methods involve an increase in data for training, due to which the sample becomes significantly larger, and the process of training it takes a very long time.

The goal of the present work is not to achieve maximum performance for the neural network. The task is to identify the features of a convolutional neural network with "averaged" learning. Therefore, it is obvious that the results can be somewhat improved, for example, due to preprocessing [21], selection of activation functions [22], and other methods. In the present work, we will consider such parameters of training a convolutional neural network as the number of epochs, the size of the mini-batch, and the generation of the initial values of the weights.

The work analyzes the effect of the size of mini-packages on the learning rate and the accuracy of pattern recognition. Optimal mini-batch sizes are evaluated. The relationship between the training accuracy and the accuracy of test samples is considered.

The change in the values of the radius vector of the scales during training is considered. Conclusions are drawn about the influence of the initial value of the balance on the recognition accuracy. A more accurate formula is obtained for the limits in which the initial values of the weights of the neural network are generated.

II. CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE

Let us consider the following convolution network architecture (Figure 1), consisting of six layers:

1) Convolutional layer. 32 cards of signs of dimension 24x24 (core bypass 5x5).

2) Subsampling layer. 32 cards of signs of dimension 12x12.

3) Convolutional layer. 64 feature cards of dimension 8x8 (core bypass 5x5).

4) Subsampling layer. 64 4x4 feature cards.

5) Fully connected layer. 512 neurons.

6) Fully connected layer. 10 neurons.

We will train the network using the method of back propagation [23] of errors, using batch gradient descent based on the handwritten digits database MNIST.

After the second, fourth and fifth layers (see Fig. 1), we use the dropout algorithm [24, 25], which is designed to

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

reduce network retraining. Dropout with probability p completely eliminates the neurons of this layer during the iteration. In our case, for the indicated layers, we choose the parameter p equal to 0.25, 0.25, and 0.5, respectively.

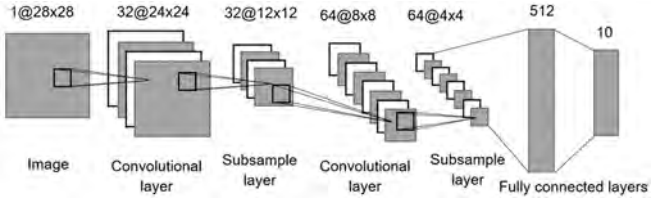


Fig. 1. The architecture of the convolutional neural network

In the first five layers, ReLU is used as an activation function [26, 27]. In the last layer, we will activate the neurons using the softmax function.

III. MINI BATCH SIZE SELECTION

We will not train the neural network throughout the sample at once, but instead in separate portions from the general sample or, in other words, mini-batches [28-31]. That is, if the training part of the MNIST database contains 60,000 images, and the size of the minibatch is 100, then the principle of the training will be as follows. We select 100 images, calculate the network error for these selected images, and obtain the changes in the current weights using the back propagation method. Next, we select the next 100 images, and in the same way we obtain the next change in the values of the weights. We will do this until we use at all the elements from the training set. In this case, it will be necessary to complete 600 learning iterations for 100 images. When all the elements of the training sample (600 iterations) are drawn through the network, we will say that one epoch has ended.

Consider the effect of the size of the mini-batch on the accuracy of recognition of digital images. To do this, we perform calculations for a different number of images in the sample. Figure 2 shows the results of calculations for several sizes of mini-batches. You may notice that the accuracy for the 16-element mini-batch quickly reaches maximum values and begins to deteriorate after 125 epochs. The accuracy for the 512 mini-batch in small epochs is low and only after 200 epochs it begins to reach maximum values.

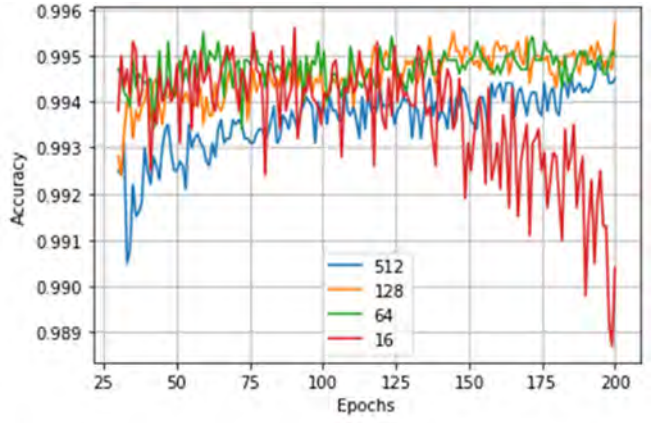


Fig. 2. Dependence of the accuracy of recognition of images from a test sample on the number of epochs with various sizes of mini-batches

For any mini-batch size from 16 to 1024, you can achieve maximum network performance, but this maximum can be achieved in a certain range of epochs. For example, for a mini-batch of size 16, the optimal range of epochs is from 25 to 75, and for a mini-batch of size 512, the optimal range starts from 250 epochs. Small sizes of mini-batches are characterized by sharp jumps in accuracy during the transition to the next epoch (the red line has sharp breaks).

Table I shows the results of computing the training time of a convolutional neural network on a CPU and on a GPU, depending on the size of mini-batches.

TABLE I. CONVOLUTIONAL NEURAL NETWORK TRAINING TIME

Mini-batch size	Training time of one epoch (seconds)		Optimal number of epochs	Total training time (seconds)	
	CPU	GPU		CPU	GPU
16	158	26	75	11850	1950
64	134	10	100	13400	1000
128	128	6	150	17920	900
512	122	4.7	200	24400	940

Thus, in our opinion, it is best to choose sizes of mini-batches from 32 to 256 elements. Such a choice ensures that in the interval from 75 to 200 epochs, the recognition accuracy on the test sample of the MNIST base will be maximum from 99.4% to 99.6%. All further considerations in this paper will be carried out with a mini-batch size equal to 128.

IV. RELATIONSHIP BETWEEN THE ACCURACY OF RECOGNITION TEST AND TRAINING SAMPLES

In the previous paragraph, it was shown that good accuracy is achieved in a wide range of epochs. In order to identify the number of epochs required for training a neural network, we will train a network with different initial weights and analyze the obtained accuracy. We will generate initial values on the segment [-*limit*, *limit*] similar to [32], where

$$limit = \sqrt{\frac{6}{N_{in}+N_{out}}}, \quad (1)$$

N_{in} is the number of input units in the vector of weights, N_{out} is the number of output units. Thus, weights between different layers are generated at different intervals. For example, for scales between the input and the first convolutional layer (see Fig. 1), the value of limit is equal to $\sqrt{6/(25+800)} \approx 0.085$. The kernel of a convolutional layer of dimension five receives 5 * 5 = 25 neurons at the input, and at the output 25 we multiply by the number of such convolution kernels 32, whence we obtain the value 800.

Fig. 3 shows graphs of the dependence of recognition accuracy in the training and test samples of MNIST on the number of epochs.

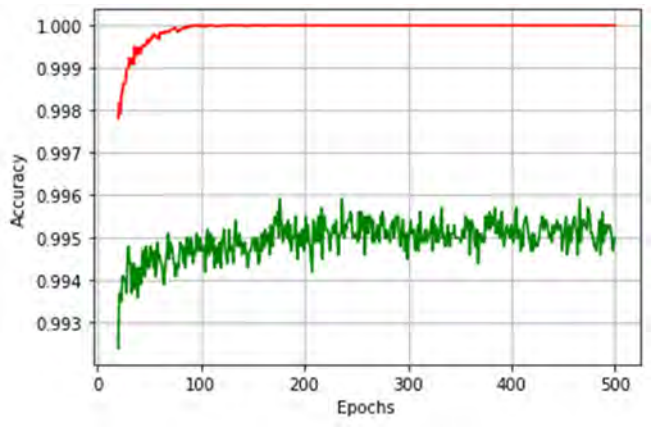


Fig. 3. Dependence of recognition accuracy on the number of epochs in the test (green broken line) and training (red broken line) samples. 500 epochs

Fig. 4 shows the results of calculations already for 1500 epochs. It can be noted that with a large number of epochs, the scatter of recognition accuracy values in the test sample (green line) decreases, but this does not lead to a decrease in recognition error.

We also note that 100% accuracy in the training set is achieved at approximately 150 epochs. In the test sample, the accuracy also remains approximately the same.

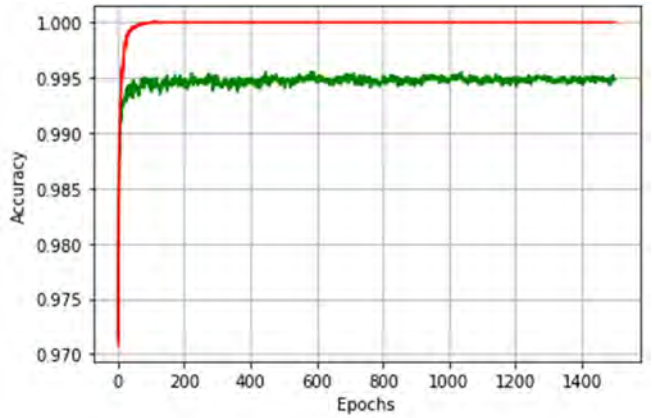


Fig. 4. The dependence of recognition accuracy on the number of epochs in the test (green broken line) and training (red broken line) samples. 1500 epochs

In order to identify the relationship between the accuracy of the test and training samples, we take 100 different models with fewer epochs (50 epochs) and draw a conclusion about the relationships between these data. The results are presented in Fig. 5.

It is seen that the maximum accuracy that the neural network was able to achieve is 99.54% in the MNIST test sample. It can also be concluded from the figure that there is no direct relationship between the accuracy of the training data set and the accuracy of the test set. In some trained neural networks, with small errors in the training set, large errors in the test set are observed.

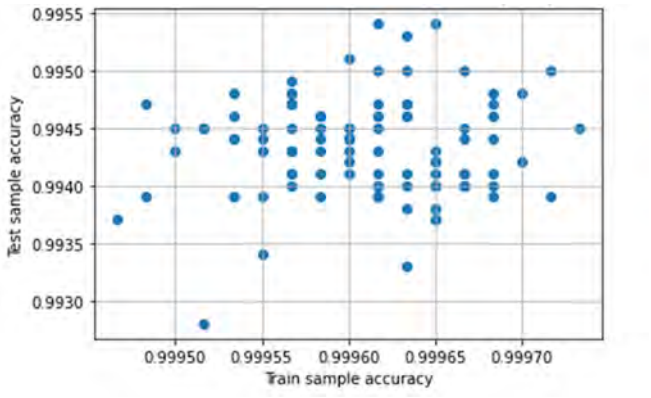


Fig. 5. The dependence of the accuracy of the test sample on the training

V. Euclidean Norm of the Weight Vector

Let us consider the Euclidean norm (the radius of the vector in the 1663370th space of weights) of the vector of the initial values of the weights. We take 10 random models and calculate the radius of each vector and the Euclidean distance between them (Table II). The radius of all weights is approximately 31.

TABLE II. Initial Euclidean Distance Between Weights of 10 Different Models

0.0	43.5	43.5	43.5	43.5	43.5	43.5	43.5	43.5	43.4
43.5	0.0	43.4	43.5	43.5	43.5	43.5	43.5	43.4	43.5
43.5	43.4	0.0	43.4	43.5	43.5	43.5	43.5	43.5	43.4
43.5	43.5	43.4	0.0	43.5	43.5	43.4	43.4	43.5	43.5
43.5	43.5	43.5	43.5	0.0	43.5	43.5	43.5	43.5	43.5
43.5	43.5	43.5	43.5	43.5	0.0	43.4	43.5	43.5	43.5
43.5	43.5	43.5	43.4	43.5	43.4	0.0	43.4	43.5	43.4
43.5	43.5	43.5	43.4	43.5	43.5	43.4	0.0	43.5	43.5
43.5	43.4	43.5	43.5	43.5	43.5	43.5	43.5	0.0	43.5
43.4	43.5	43.4	43.5	43.5	43.5	43.4	43.5	43.5	0.0

We also calculate the Euclidean distance between the vectors after learning 50 epochs (Table III). They, like the values in Table I, are obtained approximately as the same. The similarity of the obtained values is easily explained by the peculiarity of the generation of numbers. With a large dimension of vectors, we obtain pseudorandom sequences.

TABLE III. Euclidean Distance Between the Scales of 10 Different Models After Training

0.0	48.1	48.1	48.1	48.1	48.2	48.1	48.1	48.1	48.1
48.1	0.0	48.1	48.1	48.1	48.2	48.2	48.2	48.2	48.2
48.1	48.1	0.0	48.1	48.1	48.0	48.1	48.0	48.1	48.0
48.1	48.1	48.1	0.0	48.0	48.1	48.0	48.0	48.1	48.2
48.1	48.1	48.1	48.0	0.0	48.1	48.2	48.1	48.2	48.0
48.2	48.2	48.0	48.1	48.1	0.0	48.1	48.1	48.2	48.1
48.1	48.2	48.1	48.0	48.2	48.1	0.0	48.1	48.1	48.1
48.1	48.2	48.0	48.0	48.1	48.1	48.1	0.0	48.1	48.2
48.1	48.2	48.1	48.1	48.2	48.2	48.1	48.1	0.0	48.2
48.1	48.2	48.0	48.2	48.0	48.1	48.1	48.2	48.2	0.0

It should be noted that the interval of initial values selected by formula (1) corresponds to the values generated by Keras. Therefore, due to the pseudo-randomness of the

generator, we have practically the same radii of the weight vector and the same distance between the scales.

Now let us look at the behavior of the radius at large number of epochs. Fig. 6 shows the dependence of the radius of the vector on the number of epochs. Obviously, with the growth in number of epochs, the radius of the vector increases.

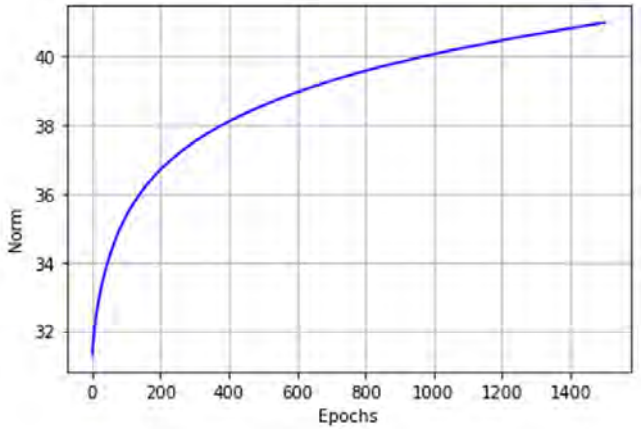


Fig. 6. The dependence of the norm (radius) on the number of epochs

In connection with this fact, a hypothesis arises that an increase in the values of the initial weights (hence, the radius of the vector) can lead to "better learning" of the neural network.

VI. Influence of the Original Scale Set on Network Accuracy

Let us change the initial weights obtained by the formula (1) by multiplying all the values of the vector of weights by a given coefficient. Next, we will train the network for the new values obtained. The results for the training sample are shown in Fig. 7.

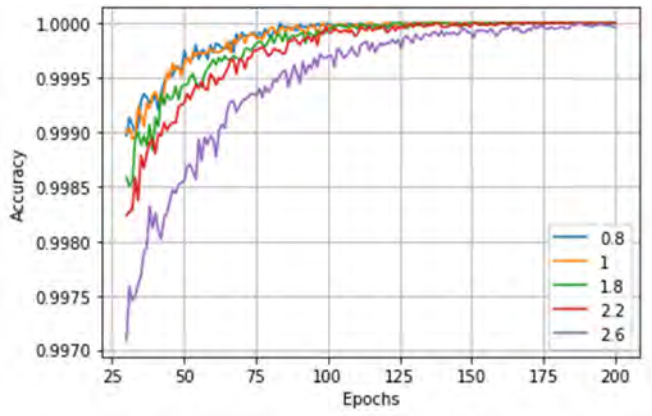


Fig. 7. Network accuracy in the training sample for different initial weights

From the analysis of the graphs in Fig. 7, one can see that the built-in option for initializing weights on the Keras framework is not optimal. Better training results (faster network training) are achieved for initial weights multiplied by a factor of 2.2.

Fig. 8 confirms our assumption. Here, weights that are of great importance, give better results.

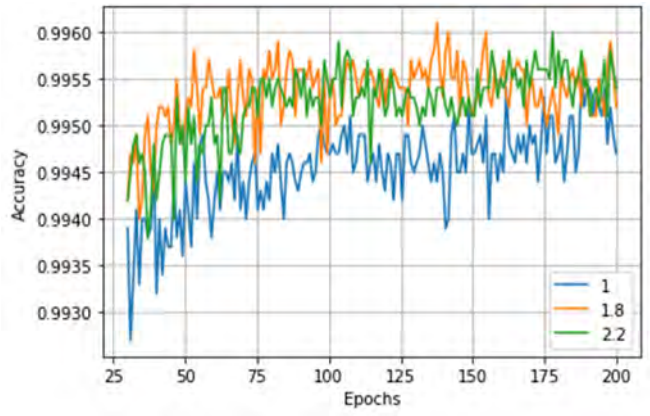


Fig. 8. The accuracy of the network in the test sample for different initial values of the weights

Thus, the following conclusion can be made. The optimal value for the interval [-*limit, limit*] is

$$limit = \sqrt{\frac{29}{N_{in}+N_{out}}}. \quad (2)$$

If the calculations are made in Keras, then after generating the initial values, one needs to multiply them by a factor of 2.2.

VII. Conclusion

A six-layer convolutional neural network is considered, which is used to recognize handwritten digit patterns from the MNIST database. The influence of the size of mini-batches on the learning speed and pattern recognition accuracy are analyzed. The optimal sizes of mini-batch are obtained. It is concluded that the optimal size of the minibatch is in the range from 32 to 256.

The relationship between the accuracy of the training sample and the accuracy of the test sample is analyzed. The change in the values of the radius vector of the scales during training is shown.

Conclusions are drawn about the influence of the initial value of the balance on the recognition accuracy. A more accurate formula is obtained for the limits, in which the initial values of the weights of the neural network are generated:

$$limit = \sqrt{\frac{29}{N_{in} + N_{out}}}.$$

The results of the work can be used to build a first-level network in hierarchical neural networks [33].

Acknowledgment

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

[1] D. N. Tumakov, D. M. Khairullina, and A. A. Valeeva, "Recovery of parameters of a homogeneous elastic layer using neural networks," Journal of Fundamental and Applied Sciences, vol. 9, 2017, pp. 1202-1220.

[2] C. Wachinger, M. Reuter, and T. Klein, "DeepNAT: Deep convolutional neural network for segmenting neuroanatomy," NeuroImage, vol. 170, 2018, pp. 434-445.

[3] R. Dautov and S. Mosin, "Technique to aggregate classes of analog fault diagnostic data based on association rule mining," Proc. of the 19th International Symposium on Quality Electronic Design, 2018, pp. 238-243.

[4] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848-6856.

[5] G. Dreyfus, "Neural Networks Methodology and Applications," Springer-Verla, 2005.

[6] L. P. J. Veelenturf, "Analysis and Applications of Artificial Neural Networks," Prentice Hall, 1995.

[7] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," Proc. of the European Conference on Computer Vision, 2014, pp. 818-833.

[8] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, 2015, pp. 85-117.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. of the IEEE, 1998, pp. 2278-2324.

[10] Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation. DeepLearning 0.1. LISA Lab.

[11] A. Krizhevsky, "ImageNet: Classification with deep convolutional neural networks," Proc. of the 25th International Conference on Neural Information Processing Systems, 2012, pp. 1097-1105.

[12] S. Neha, J. Vibhor, and M. Anju, "An analysis of convolutional neural networks for image classification," Procedia Computer Science, vol. 132, 2018, pp. 377-384.

[13] P. I. Earnest and M. C. Krishna, (2016) "Human action recognition using genetic algorithms and convolutional neural networks," Pattern, vol. 59, 2016, pp. 199-212.

[14] H. Xuanyu and Z. Wei, "Emotion recognition by assisted learning with convolutional neural networks," Neurocomputing, vol. 291, 2018, pp. 187-194.

[15] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, and M. Daoudi, "Lip reading with hahn convolutional neural networks," Image and Vision Computing, vol. 88, 2019, pp. 76-83.

[16] D. Tome, F. Monti, L. Baroffio, L. Bondi, and S. Tubaro, "Deep convolutional neural networks for pedestrian detection," Signal Processing: Image Communication, vol. 47, 2016, pp. 482-489.

[17] The MNIST database handwritten digits. - URL: http://yann.lecun.com/exdb/mnist.

[18] P. Simard, S. Dave, and J. Platt, (2003) "Best practices for convolutional neural networks applied to visual document analysis," Proc. of the Seventh International Conference, 2003, pp. 958-962.

[19] D. Cires, M. Ueli, and S. Jürgen, "Multi-column deep neural networks for image classification," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3642-3649.

[20] V. Romanuke, "Training data expansion and boosting of convolutional neural networks for reducing the MNIST dataset error rate," Research Bulletin of NTUU "Kyiv Polytechnic Institute", vol. 6, 2016, pp. 29-34.

[21] S. Mosin, "Machine learning and data mining methods in testing and diagnostics of analog and mixed-signal integrated circuits: Case study," Communications in Computer and Information Science, vol. 968, 2019, pp. 240-255.

[22] R. Latypova and D. Tumakov, "Method of selecting an optimal activation function in perceptron for recognition of simple objects," Proc. of the 16th IEEE East-West Design and Test Symposium, 2018, pp. 390-394.

[23] R. Hecht-Nielsen, "Theory of the backpropagation neural network," Neural Networks, vol. 2, 1992, pp. 65-93.

[24] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, 2012.

[25] P. Baldi and P. Sadowski, "Understanding dropout," Proc. of the 26th International Conference on Neural Information Processing Systems, 2013, pp. 2814-2822.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines" Proc. of the 27th International Conference on International Conference on Machine Learning, 2010, pp. 807-814.

[27] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," Proc. of the 33rd International Conference on International Conference on Machine Learning, 2016, pp. 2217-2225.

[28] P. M. Radiuk, "Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets," Information Technology and Management Science, vol. 20(1), 2017, pp. 20-24.

[29] Y. Zhang, H. Qu, C. Chen, and D. Metaxas, "Taming the Noisy Gradient: Train Deep Neural Networks with Small Batch Sizes," Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), 2019, pp. 4348-4354.

[30] C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl, "Measuring the effects of data parallelism on neural network training," Journal of Machine Learning Research, vol. 20, 2019, pp. 1-49.

[31] N. Z. T. Abdulnabi and O. Altun, "Batch size for training convolutional neural networks for sentence classification," Journal of Advances in Technology and Engineering Research , vol. 2, 2016, pp.156-163.

[32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Proc. of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249-256.

[33] Z. Kayumov, D. Tumakov, and S. Mosin, "Hierarchical convolutional neural network for handwritten digits recognition," Procedia Computer Science, vol. 171, 2020, pp. 1927-1934.

Designing a Single-Band Monopole Six-Tooth-Shaped Antenna with Preset Matching

Angelina Markina
Institute of Computantional Mathematics and Information Technologies
Kazan Federal University
Kazan, Russia
m8angelina@gmail.com

Dmitrii Tumakov
Institute of Computantional Mathematics and Information Technologies
Kazan Federal University
Kazan, Russia
dtumakov@kpfu.ru

***Abstract*— The design of a symmetrical six-tooth-shaped monopole microstrip antenna is considered. The effect of rectangular cutouts on the radiator and the length and the width of the radiator on the reflection coefficient of the base frequency antenna is studied. A nonlinear regression model with good accuracy is constructed for this characteristic. An approach to design using regression models of a tooth-shaped antenna for the desired wireless network parameters is presented. Optimization problems are constructed and numerically solved, which allows to quickly determine the optimal values of the geometric parameters of the tooth-shaped radiator. The application of this approach to the design of a six-tooth-shaped single-band antenna for Wi-Fi applications is demonstrated.**

Keywords—antenna design, six-tooth-shaped antenna, monopole microstrip antenna, single-band Wi-Fi antenna, reflection coefficient, regression models.

I. INTRODUCTION

Monopole microstrip antennas are widely used in wireless systems [1-3]. Such antennas have compact dimensions, light weight, simple manufacturing technology and support a wide range of frequencies for various applications [4, 5], including Wi-Fi applications.

The antenna reflector is usually a flat metal plate of a certain shape, and by changing the geometry of the reflector in such antennas, the desired electrodynamic characteristics are attained [6, 7]. Rectangular reflectors of various shapes are often used such as M-shaped [8], U-shaped [9], lamp-shaped radiating patch [10] and many others [11, 12].

Engineers often do not know with which form of the reflector to begin designing the antenna and how to further change its shape in order to obtain an antenna with the optimal desired electrodynamic characteristics. Therefore, designing a well-matched antenna is a rather lengthy and time-consuming process.

To quickly improve the matching of the microstrip antenna at a frequency of 2.45 GHz, the authors of [13] used the knowledge-based neural network. To improve the electrodynamic characteristics of the antenna, genetic algorithms are also used [14, 15] and the defected ground structure technique [16] is used.

To speed up the process of designing microstrip antennas, we propose using regression models that describe the relationship between the characteristics of the antenna and its geometric parameters [17]. First, one must first conduct a study over a certain family of antennas, for example, tooth-shaped antennas [18]. Then a connection between the electrodynamic characteristics of the antenna and the reflector geometry in the form of mathematical models must be established, and optimization problems for these antenna characteristics must be formulated. For example, in [19-21] single- and double-band symmetric four-tooth-shaped microstrip antennas were designed using regression models and problems of optimizing the electrodynamic characteristics of the antenna were solved.

In the present work, a change in the values of the reflection coefficient with varying the basic geometric parameters of the reflector is studied. A nonlinear regression model is constructed for the reflection coefficient, depending on the width and length of the reflector and the size of six symmetrical cutouts. The design process of a single-band microstrip antenna at a frequency of 2.44 GHz is presented as a solution to optimization problems with respect to the basic electrodynamic characteristics of the antenna. During the design, two well-matched antennas at a frequency of 2.43 GHz were obtained.

II. PROBLEM STATMENT

The purpose of this article is to design an omnidirectional monopole microstrip antenna at a frequency of 2.44 GHz with the preset matching for Wi-Fi applications. In this case, the antenna should have the largest possible bandwidth at the desired frequency.

Let us consider the antenna design presented in Figure 1. Its radiator has the correct symmetric six-tooth-shaped form and is described by the parameters a_R, b_R, d_R, c_R, where a_R and b_R are the width and length, d_R is the depth of symmetrical cutouts, c_R is the length of the ridges calculated as the ratio of the length of the radiator to the sum of the number of cutouts and the number of teeth on one side of the radiator.

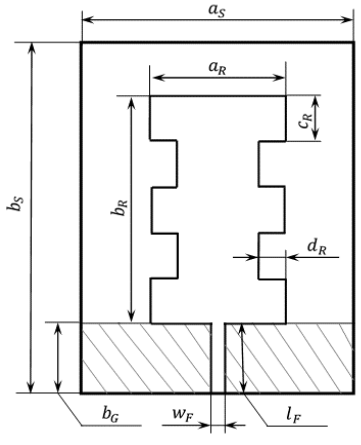


Fig. 1. The construction of the symmetrical six-tooth-shaped microstrip antenna

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

This shape of the radiator is obtained from a rectangular radiator by adding three symmetrical cutouts on its long side. The radiator is located on the front side of the substrate and is fed from a source with a resistance of 50 Ω through a power line with the length l_F and the width w_F.

On the reverse side of the substrate, there is a rectangular metal plate – ground plane (shaded area in Fig. 1). Ground plane has the width equal to the width of the substrate and the length b_G equal to the length of the power line, i.e. $b_G = l_F$. The substrate is uniformly filled with a dielectric with a dielectric constant $\varepsilon_r = 4.5$ with a material density ρ = 1000 kg / m^3 and a dielectric loss tangent tan δ = 0. The dimensions of the substrate are described by the parameters a, b_S and t_S, where a_S and b_S are the width and the length of the substrate, t_S is the thickness. The dimensions of the parameters of the substrate and the feeding line are presented in Table I.

TABLE I. VALUES OF ANTENNA PARAMETERS

Antenna parameters	a_S	b_S	t_S	l_F	w_F	b_G
Values of parameters, mm	30	75	1	15	1	15

Let us determine the values of the radiator parameters a_R, b_R and d_R, to which the six-tooth-shaped antenna will be tuned for the Wi-Fi resonant frequency (2.44 GHz); which will have good matching and a wide bandwidth.

III. INVESTIGATION OF THE INFLUENCE OF THE LENGTH, WIDTH OF THE RADIAOTR AND DEPTH OF CUTOUTS ON THE REFLECTION COEFFICIENT

Let us study the dependence of the reflection coefficient S_{11} on the width a_R and the length b_R of the radiator, as well as on the depth of six symmetrical rectangular cutouts d_R.

To do this, we consider three sizes of the length of the radiator $b_R = 24$, 32.5 and 41 mm; we set the parameter c_R from the relation $b_R/5$. In the FEKO program, we carry out numerical experiments, in which we vary the parameter values with a step of 0.5 mm

- d_R = 0.4…4.9 mm at a_R = 10 mm.
- d_R = 0.4…7.4 mm at a_R = 15 mm.
- d_R = 0.4…9.9 mm at a_R = 20 mm.
- d_R = 0.4…11.9 mm at a_R = 24 mm.

It should be noted that the calculation of the electrodynamic characteristics of one tooth-shaped microstrip antenna in the FEKO takes approximately one hour. The total number of calculated tooth-shaped antennas is 207. Thus, the time taken to prepare the data for the study is approximately 230 hours.

A. The Influence of the Radiator Length and the Depth of Cutouts

Let us consider in Fig. 2 changes in the reflection coefficient S_{11} when varying the size of the depth of symmetrical cutouts on the radiator and three fixed values of the length of the radiator with $a_R = 10$ mm. Here, the blue graph with round markers corresponds to a radiator of the length $b_R = 24$ mm, the orange graph with square markers corresponds to a radiator of $b_R = 32.5$ mm, the green graph with triangular markers corresponds to a radiator of the length $b_R = 41$ mm.

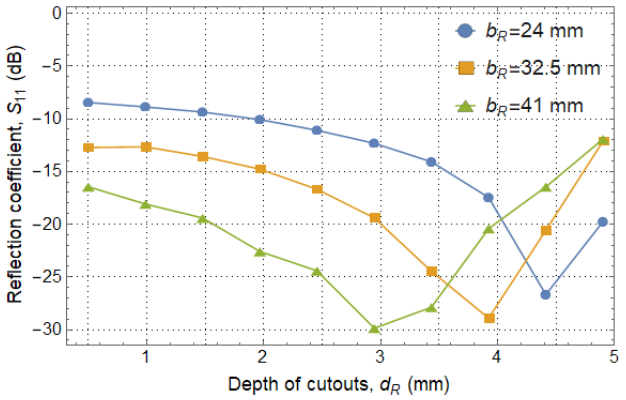


Fig. 2. The dependences of the reflection coefficient S_{11} on the depth of the cutouts d_R with various radiator length b_R at the radiator width $a_R = 10$ mm

We note that in this figure we observe a similar change in the values of the reflection coefficient with increasing depth of cutouts, i.e. on all graphs, S_{11} decreases and reaches its lowest value. For the radiator with $b_R = 24$ mm and $d_R = 4.4$ mm, the minimum value $S_{11} = -26.67$ dB; for the radiator with $b_R = 32.5$ mm and $d_R = 3.9$ mm, we have $S_{11} = -28.90$ dB, with $b_R = 32.5$ mm and $d_R = 2.9$ mm; the minimum value is $S_{11} = -29.86$ dB. With a further increase in the depth of cutouts, we see that the values of the reflection coefficient increase. Note that at the maximum depth of cutouts $d_R = 4.9$ mm and for radiator lengths 32.5 mm and 41 mm, the value of the S_{11} is slightly larger than that with small cutouts $d_R = 0.4$ mm.

Next, we consider in Fig. 3 changes in the reflection coefficient S_{11} at $a_R = 20$mm. Here, the behavior of marked graphs S_{11} differs markedly from the behavior of graphs in Fig. 2. The values of the reflection coefficient for different radiator lengths do not differ much; they also decrease with increasing depth of cutouts and reach a minimum value with a maximum depth of cutouts $d_R = 9.9$ mm. Note that for the radiator $b_R = 24$ mm the minimum value is $S_{11} = -20.79$ dB, for $b_R = 32.5$ mm we have $S_{11} = -25.82$ dB and for $b_R = 41$ mm the minimum value is $S_{11} = -28.96$ dB.

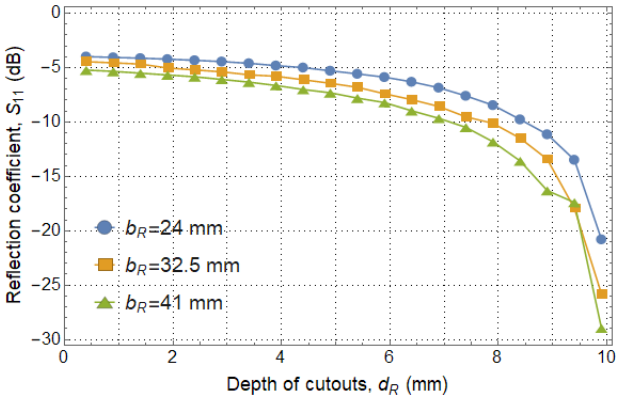


Fig. 3. The dependences of the reflection coefficient S_{11} on the depth of the cutouts d_R with various radiator length b_R at the radiator width $a_R = 20$ mm

Consequently, a decrease in the size of symmetrical cutouts leads to a decrease in the reflection coefficient, and slightly elongated radiators ($b_R/a_R < 2$) with a cutout depth d_R close to $a_R/2$ have the smallest values of the S_{11}. For $b_R/a_R > 2$ (elongated radiators), the minimum value of the reflection coefficient reaches at a value of d_R close to a complex ratio of the parameters a_R, b_R, d_R.

B. The Influence of Radiator Width and Depth of Cutouts

Now we present in Fig. 4 a change in the reflection coefficient for a fixed radiator length $b_R = 41$ mm and its different widths. Here, the blue graph with round markers corresponds to a radiator width $a_R = 10$ mm, the orange graph with square markers corresponds to a radiator $a_R = 15$ mm, the green graph with triangular markers corresponds to a radiator $a_R = 20$ mm, the pink graph with white round markers corresponds to a radiator $a_R = 24$ mm.

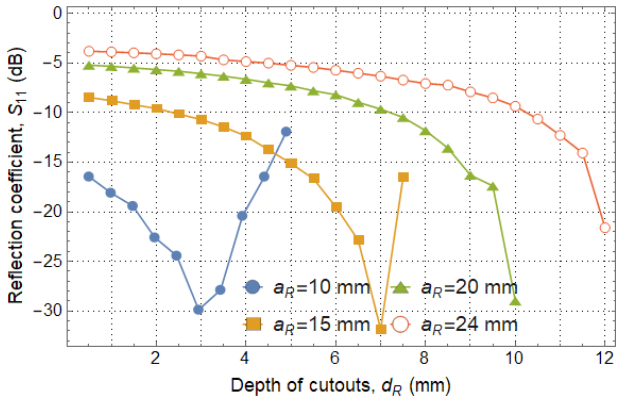


Fig. 4. The dependences of the reflection coefficient S_{11} on the depth of the cutouts d_R with various radiator width a_R at the radiator length $b_R = 41$ mm

In this case, we see an improvement in antenna matching with increasing size of the cutouts d_R. Note that the smaller the difference $(a_R - 2d_R)$, the smaller the values of S_{11} that the radiator with the ratio of length to width $b_R/a_R < 2$ (not very elongated radiators) has.

Let us compare the results according to the graphs in Fig. 2-4 and draw the following conclusions. With a small cutout size d_R from 0.5 to 4.5 mm for radiator with $b_R/a_R < 2$, the reflection coefficient S_{11} has closer values for different b_R, than for radiator with $b_R/a_R > 2$. Therefore, in order to obtain good agreement for slightly elongated radiator, it is better to use a larger cutout size, and for strongly elongated radiator, it is better to use average values of d_R. Thus, by varying the ratio of length to width and the difference $(a_R - 2d_R)$, it is possible to obtain an antenna with well-matching.

C. Constructing of a regression model

We use the above conclusions and proceed to the construction of a regression model for the reflection coefficient at the base frequency. Note that the graphs of values S_{11} form a parabola in the case of strongly elongated radiators $(b_R/a_R > 2)$ and have the form of a power function for the shape of the radiator close to square $(b_R/a_R < 2)$. It is also necessary to take into account the sharp decrease in the values of S_{11} when the values of d_R are close to $a_R/2$ and the ratio of the length to width of the radiator.

We construct the regression model of S_{11} for the base frequency in the following form:

$$\bar{S}_{11}(a_R, b_R, d_R) = c_1 \left(\frac{a_R}{b_R}\right)^{c_2 \frac{b_R}{a_R}} b_R + \frac{c_3 d \sqrt{a_R - d_R}}{a_R^5} + c_4 \frac{c_4 (a_R - 2\,d_R)^2}{b_R} + c_5 \frac{a_R}{a_R - d_R} + c_6 a_R, \quad (1)$$

where S_{11} is measured in dB, radiator length b_R, radiator width a_R and depth of cutouts d_R are measured in mm, and the coefficients c_n $(n = \overline{1,6})$ are assumed to be unknown.

Using the least squares method, we determine the unknown coefficients in (1) and obtain $c_1 = 16.5901$, $c_2 = 0.2100$, $c_3 = 5486.0326$, $c_4 = -0.2484$, $c_5 = -19.7009$, $c_6 = 0.2425$. We use the formulas from [22], and the mean square error:

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(S_{11}^i - \bar{S}_{11}^i\right)^2}$$

and relative absolute error [22]:

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{S_{11}^i - \bar{S}_{11}^i}{S_{11}^i}\right| \cdot 100\%,$$

where S_{11}^i are the known values, $\bar{S}_{11}^i$ are the values calculated by the formula (1). Then we obtain that the regression model (1) is built with very good accuracy and has errors: $\epsilon \approx 2.85$ dB, $\sigma \approx 10.6\%$. Therefore, formula (1) can be used to find the minimum values S_{11} at given constrains on the parameters of the radiator.

IV. Designing a Single-Band Antenna

We proceed to the design of a six-tooth-shaped single-band antenna at the frequency of 2.44 GHz and find the values of the radiator parameters a_R, b_R and d_R.

A. First Method

We use the approach presented in our work [21], but in a different way: we define the regression model (1) for the optimization problem for the reflection coefficient S_{11} without further improving the electrodynamic characteristics and minimizing S_{11} using the gradient descent method.

1) Determination of the radiator length at the frequency of 2.7 GHz.

First, we determine the length of the rectangular radiator. It was revealed in [19] that the base frequency strongly depends on the length of a rectangular and tooth-shaped radiator. We calculate b_R at the frequency of 2.7 GHz, slightly offset from 2.44 GHz, since in the next step we adjust the base frequency by adding symmetrical cutouts.

We use the quadratic function (2) obtained in [19] for a rectangular radiator at the base frequency:

$$f_1(b_R) = 5.18 - 1.1\, b_R + 0.083 b_R^2. \quad (2)$$

We substitute the value $f_1 = 2.7$ GHz in (2) and determine b_R = 28 mm.

2) Determination of raditor width and depth of cutouts at the frequency of 2.44 GHz

Next, we add symmetrically two rectangular cutouts on the long sides of the rectangular radiator and obtain a symmetrical six-tooth-shaped antenna. Let us take the length of the radiator a little longer than calculated in step 1, for example, 30 mm, since the depth of the cutouts reduces the values of the base frequency. We formulate two optimization problems: in the first problem, we minimize the objective function (3), presented as the sum of the squares of the difference of the

required values for the bandwidth and the base frequency with weights α and β, at the linear constraints (4):

$$\alpha\,(BW(a_R, b_R, d_R) - 1)^2 + \beta\,(f(a_R, b_R, d_R) - 2.44)^2 \to \min, \quad (3)$$

$$\begin{cases} 0.5 \text{ mm} \le d_R \le \dfrac{a_R}{2} - 0.5 \text{ mm}, \\ 5 \text{ mm} \le a_R \le \dfrac{b_R}{3} \text{ mm}, \\ S_{11}(a_R, b_R, d_R) \le -15 \text{ dB}. \end{cases} \quad (4)$$

In the second problem, we maximize the bandwidth under the same restrictions (4):

$$BW(a_R, b_R, d_R) \to \max, \quad (5)$$

where $f_1(a_R, b_R, d_R)$ is the regression model of the base frequency constructed in [23], $BW(a_R, b_R, d_R)$ is the regression model of the bandwidth constructed in [24], $S_{11}(a_R, b_R, d_R)$ is regression model (1) for the reflection coefficient. In constraints (4), the first condition ensures that the cutouts on the sides will not intersect inside the radiator, the second condition limits the width of the radiator and allows one to obtain an elongated tooth-shaped form; the third condition means that the antenna must be well matched and have a reflection coefficient of not more than –15 dB.

In the objective function (3), weights α and β are set to 1 and 0.8, respectively. In (3) - (5) we substitute the length of the radiator 30 mm and numerically find the solution to the optimization problems (3) - (4) and (5) - (4) in the Wolfram Mathematica package using the built-in Minimize functions *Minimize*[{*f*, *cons*}, {*x*, *y*, …}] and Maximize [{*f*, *cons*}, {*x*, *y*, …}], respectively. Thus, we obtain the parameters a_R and d_R, for which the objective functions (3) and (5) achieve optimal values, taking into account restrictions (4).

As a result, we obtain two antennas with a six-tooth-shaped radiator with parameters:

- $a_R = 7.34$ mm and $d_R = 0.62$ mm, the minimum of the objective function (3) is 0.017 GHz.
- $a_R = 7.16$ mm and $d_R = 0.5$ mm from problem (5)-(4), the maximum of the objective function (5) is 0.895 GHz.

The antennas simulation and a numerical calculation are performed in the FEKO program. The results of calculating the reflection coefficient are presented in Fig. 5.

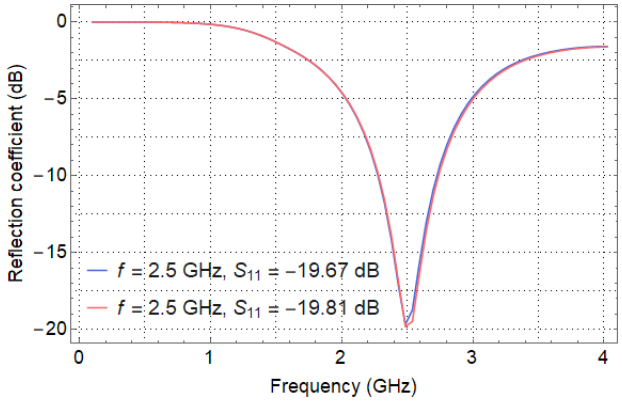


Fig. 5. The dependence of the reflection coefficient on the frequency.

The blue line corresponds to the radiator with $a_R = 7.34$ mm, $d_R = 0.62$ mm; the red line corresponds to the radiator with $a_R = 7.16$ mm, $d_R = 0.5$ mm.

It can be seen that the electrodynamic characteristics of the antennas almost coincide: the first resonance is at the frequency of 2.5 GHz and the reflection coefficients are -19.67 dB and -19.81 dB. This is due to the fact that the dimensions of the geometric parameters of the two radiators are close to each other.

B. Second Method

We now formulate optimization problems in such a way that in one step utilizing regression models one can immediately find three radiator parameters a_R, b_R and d_R. The first optimization task has the form:

$$\alpha\,(BW(a_R, b_R, d_R) - 1)^2 + \beta\,(f(a_R, b_R, d_R) - 2.44)^2 \to min \quad (6)$$

$$\begin{cases} 0.5 \text{ mm} \le d_R \le \dfrac{a_R}{2} - 0.5 \text{ mm}, \\ 5 \text{ mm} \le a_R \le \dfrac{b_R}{2} \text{ mm}, \\ 10 \text{ mm} \le b_R \le 30 \text{ mm}, \\ S_{11}(a_R, b_R, d_R) \le -20 \text{ dB}. \end{cases} \quad (7)$$

The objective function (6) remains unchanged. In constraints (7), the conditions changed for the reflection coefficient and the width of the radiator, and the condition for limiting the length of the radiator is added.

The second optimization problem has the form:

$$BW(a_R, b_R, d_R) \to max, \quad (8)$$

$$\begin{cases} 0.5 \text{ mm} \le d_R \le \dfrac{a_R}{2} - 0.5 \text{ mm}, \\ 5 \text{ mm} \le a_R \le \dfrac{b_R}{2} \text{ mm}, \\ 10 \text{ mm} \le b_R \le 70 \text{ mm} \\ S_{11}(a_R, b_R, d_R) \le -15 \text{ dB}, \\ 2.42 \text{ GHz} \le |f(a_R, b_R, d_R)| \le 2.48 \text{ GHz}. \end{cases} \quad (9)$$

Here, the objective function (6) also remains unchanged. In constraints (9), the condition changed for the width of the radiator, and conditions were added to limit the length of the radiator as well as to limit the values of the base frequency.

In the objective function (6), weights α and β are set to 1 and 0.5, respectively, and numerical solution to the optimization problems (6)-(7) and (8)-(9) are obtained in the Wolfram Mathematica package. Thus, for each of the task accordingly, the following parameters of the radiator are obtained:

- $a_R = 5.49$ mm, $b_R = 30$ mm and $d_R = 1.20$ mm, the minimum of the objective function (6) is equal to 0.093 GHz.
- $a_R = 7.39$ mm, $b_R = 31.22$ mm and $d_R = 0.5$ mm, the maximum of the objective function (8) is equal to 0.892 GHz.

A numerical calculation of the electrodynamic characteristics of the designed antennas is performed in the

FEKO program. The results of the calculations are presented in Fig. 6. The blue line corresponds to the radiator with $a_R = 5.49$ mm, $b_R = 30$ mm and $d_R = 1.20$ mm; the red line shows the radiator with $a_R = 7.39$ mm, $b_R = 31.22$ mm and $d_R = 0.5$ mm.

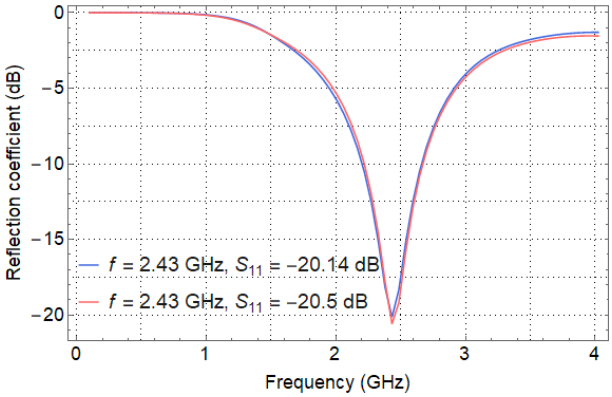


Fig. 6. The dependence of the reflection coefficient on the frequency.

According to Fig. 6, the electrodynamic characteristics of the antennas almost coincide: the first resonance is at the frequency of 2.43 GHz and the reflection coefficients are –20.14 dB and –20.5 dB. However, the dimensions of the geometric parameters of the two radiators are different. The coincidence of characteristics can be explained by the fact that the difference in the sizes of the same parameters of the radiator is compensated by the difference in the values of the opposite parameters.

Fig. 7 presents a radiation pattern at the frequency of 2.43 GHz for a six-tooth-shaped antenna with radiator parameters $a_R = 5.49$ mm, $b_R = 30$ mm and $d_R = 1.20$ mm. It can be seen that the antenna is omnidirectional with a maximum gain of 1.75.

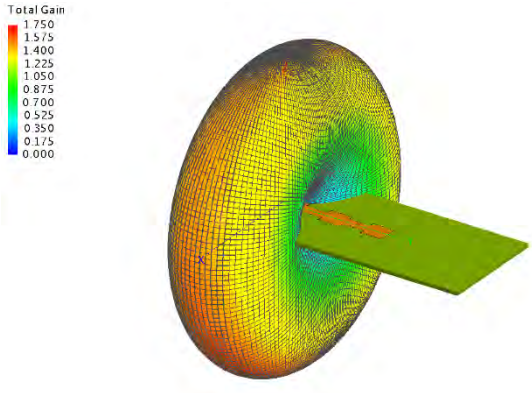


Fig. 7. The radiation pattern at 2.43 GHz

Comparing the two approaches, we can draw the following conclusions. The design of six-tooth-shaped antennas, taken as a solution to the optimization problem with regression models, is very fast. By passing the long sequential selection of suitable sizes of several parameters of the radiator, it is possible to efficiently quickly build an antenna that has the required characteristics.

V. Conclusion

The work considers a microstrip antenna with a symmetrical six-tooth-shaped radiator. The dependence of the reflection coefficient of the base frequency on the size of the teeth, the length and width of the radiator are studied. The changes in the values of the reflection coefficient with varying depths of rectangular cutouts and the dimensions of the radiator are graphically shown. An expression for the reflection coefficient as a function of the geometric parameters of the radiator is obtained.

An approach to design using regression models of a six-tooth-shaped antenna for the desired electrodynamic parameters is presented. As a result of the numerical solution of optimization problems, the optimal geometric parameters of the radiator are determined, giving a well-matched antenna at the frequency of 2.43 GHz.

It should be noted that the proposed approaches can be used to improve the electrodynamic properties of the antenna for certain parameters of a wireless communication system, including for Wi-Fi applications.

Acknowledgment

This work was supported by the research grant of Kazan Federal University. The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

[1] A.Q. Khan, M. Riaz, and A. Bilal, “Various types of antenna with respect to their applications: a review,” International Journal of Multidisciplinary Sciences and Engineering, vol. 7, no. 3, 2016, pp. 1–8.

[2] C.A. Balanis, Antenna Theory: Analysis and Design, John Wiley & Sons, Inc., Hoboken, New Jersey, 2016.

[3] D. Guha and Y.M.M. Antar, Microstrip and Printed Antennas: New Trends, Techniques and Applications, John Wiley & Sons, 2011.

[4] M.P. Joshi, V.J. Gond, “Microstrip Patch Antennas for Wireless Communication: A Review,” Proc. on ICEI 2017, 2017, pp. 96-99.

[5] N. Sharma, S.S. Bhatia, V. Sharma and J.S. Sivia, “An Octagonal Shaped Monopole Antenna for UWB Applications with Band Notch Characteristics,” Wireless Personal Communications, vol. 111, no. 3, 2020, pp. 1977–1997.

[6] J.R. Panda, A. S. R. Saladi, and R. S. Kshetrimayum, “A compact printed monopole antenna for dual-band RFID andWLAN applications,” Radioengineering 20, 2011, pp. 464–467.

[7] H. F. Huang and Y. Hu, “A compact dual-band printed monopole antenna for WiMAX/WLAN applications,” Progress InElectromagnetics Research, vol. 49, 2014, pp. 91–97.

[8] B. Azarm, J. Nourinia, C. Ghobadi, and M. Karamirad, “Novel design of dual band-notched rectangular monopoleantenna with bandwidth enhancement for UWB applications,” in Electrical Engineering (ICEE), Iranian Conference on(IEEE, 2018), 2018, pp. 567–571.

[9] J. Ghimire and D. Y. Choi, “Design of a compact ultrawideband U-shaped slot etched on a circular patch antenna withnotch band characteristics for ultrawideband applications,” International Journal of Antennas and Propagation, vol. 2019, 2019.

[10] S. Yadav, A. K. Gautam, and B. K. Kanaujia, “Design of dual band-notched lamp-shaped antenna with UWB characteris-tics,” International Journal of Microwave and Wireless Technologies, vol. 9, 2015, pp. 395–402.

[11] M.P. Joshi and V.J. Gond, “Microstrip patch antennas for wireless communication: A review,” in 2017 InternationalConference on Trends in Electronics and Informatics, 2017, pp. 96–99.

[12] M. El-Sayed, N. Gad, M. El-Aasser, and A. Yahia, “Slotted rectangular microstrip-antenna design for radar and 5Gapplications,” in 2020 International Conference on Innovative Trends in Communication and Computer Engineering, 2020, pp. 330–334.

[13] Y. Chen, Y.B. Tian, Z. Qiang, and L. Xu, “Optimisation of reflection coefficient of microstrip antennas based on KBNN exploiting GPR model”, IET Microwaves, Antennas & Propagation, vol. 12, no. 4, 2017, pp. 602-606.

[14] J.S. Smith and M.E. Baginski, “Thin-wire antenna design using a novel branching scheme and genetic algorithm optimization,” IEEE

Transactions on Antennas and Propagation, vol. 67, no. 5, 2019, pp. 2934-2941.

[15] Z. Zhang, S. Yang, M. Liu, S. Deng, and L. Li, "Design of an UWB microstrip antenna with DGS based on genetic algorithm," 2019 21st International Conference on Advanced Communication Technolog, IEEE, 2019, pp. 228-232.

[16] M.K. Khandelwal, B.K. Kanaujia, and S. Kumar, "Defected ground structure: fundamentals, analysis, and applications in modern wireless trends," International Journal of Antennas and Propagation, vol. 2017, 2017, Article ID 2018527.

[17] D.N. Tumakov, G.V. Abgaryan, D.E. Chickrin, and P.A. Kokunin, "Modeling of the Koch-type wire dipole," Applied Mathematical Modelling, vol. 51, 2017, pp. 341-360.

[18] A.G. Markina, N.B. Pleshchinskii, and D.N. Tumakov, "On electrical characteristics of comb-shaped microstrip antennas," Young Researchers in Electrical and Electronic Engineering (EIConRus), 2017 IEEE Conference of Russian, IEEE, 2017, pp. 179-183.

[19] A. Markina, D. Tumakov, and N. Pleshchinskii, "Designing a symmetrical eight-teeth-shaped microstrip antenna for Wi-Fi applications," Proc. of 16th EWDTS, 2018, pp. 491–495.

[20] D.N. Tumakov, A.G. Markina, and I.B. Badriev, "Fast method for designing a well-matched symmetrical four-tooth-shaped microstrip antenna for Wi-Fi applications," Journal of Physics: Conference Series, vol. 1158, no. 4, 2019, 042029.

[21] A. Markina and D. Tumakov, "Designing the Four-Tooth-Shaped Microstrip Antenna for Wi-Fi Applications," 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA), IEEE, 2019, pp. 25-30.

[22] J.O. Rawlings, S.G. Pantula, and D.A. Dickey, Applied regression analysis: a research tool, Springer Science & Business Media, 2001.

[23] A. Markina, D. Tumakov, and N. Pleshchinskii, "On base frequency of the symmetrical six-tooth-shaped microstrip antenna," National Academy of Managerial Staff of Culture and Arts Herald, no. 3, 2018, pp. 983-989.

[24] A. Markina and D. Tumakov, "On bandwidth of the symmetrical six-tooth-shaped monopole microstrip antenna," Bioscience Biotechnology Research Communications, vol. 12, no. 5, 2019, pp. 264-271.

Hybrid implementation of Twofish, AES, ElGamal and RSA cryptosystems

Elza Jintcharadze
Department of Computer Science, Batumi Shota Rustaveli State University
Batumi, Georgia
elza.jintcharadze@bsu.edu.ge

Maksim Iavich
Cyber Security Department, School of technology Caucasus University
Tbilisi, Georgia
m.iavich@scsa.ge

***Abstract* — Nowadays, for achieving information security and to provide security against unauthorized access cryptography plays an important role. To ensure high-security level there are different types of cryptographic methods. This paper presents implementation and analysis of new hybrid cryptosystems. Main objectives of this paper are to emphasize on better performance, maximum speed of an algorithm, checking effectiveness and comparison with other algorithms. In the paper is proposed two new hybrid algorithms using combination of both symmetric and asymmetric cryptographic algorithm such as Twofish, AES, RSA and ElGamal. To analyze results was used JAVA program implementation. The results shows that the proposed hybrid algorithm AES+RSA is significantly secure. However, Twofish + RSA hybrid has other advantages like better computation time, the size of cipher text, and the memory consumption.**

***Keywords*—** *Hybrid encryption; Symmetric cryptography; Asymmetric cryptography; Ciphertext; Plaintext; Cryptosystems comparison.*

I. TWOFISH

Twofish algorithm is a one of the popular symmetrical algorithm. This algorithm belongs to block type algorithm family. It has 128 bits block size. Key size of this algorithm is variable and changes to 256 bits. Predecessor of Twofish algorithm is Blowfish algorithm [11]. The Twofish and Blowfish algorithms has quite similar structure. Twofish algorithm's main feature is that it has complex scheme of encryption and S-blocks are pre-calculated, also key-dependent. The n-bit keys are divided into two parts: 1. half of keys are used as the encryption key; 2. another half of keys are used to modify algorithm. In terms of speed Twofish can show better performance than AES. To compare Twofish with AES cryptosystem, both algorithms can support different key sizes (from 128, 192, and 256 bits), therefore their resistant against brute force attack are equal.

The main difference between AES and Twofish ciphers is that for data encrypting AES uses a substitution-permutation network. Feistel network is used to perform data encryption by Twofish. Apparently Twofish algorithm is more complicated compared to other older standards like DES (Data Encryption Standard) and 3DES (Triple DES).

II. AES - ADVANCED ENCRYPTION STANDARD

AES is a symmetric, block cipher type algorithm. AES has three blocks each with a block size of 128, 192 and 256 bits [18, 19]. AES uses different length keys to encrypt and decrypt data. For example, AES-128 uses 128 bit length key. For the reason of different key length, during encryption AES needs the various number of rounds. If AES key sizes are128 bits, 192 bits, and 256 bits, respectively the number of rounds will be 10, 12, and 14.

On the other hand, while using Twofish algorithm for any size of key the number of rounds are not variable. Number of rounds for Twofish algorithm is 16.

To compete different algorithms we can consider their resistance against attacks. Throughout competition security and vulnerability against attacks are the most vital factors. Implementation and suitability are also one important criteria while comparison of cryptography algorithms.

When it comes to hardware requirements, AES is very efficient than Twofish. Although, to encrypt data AES requires less memory and fewer cycles.

III. ELGAMAL

ElGamal encryption system is a public-key cryptosystem. The security of this algorithm depends on the difficulty of finding discrete logarithm. Using ElGamal algorithm simple data can be encrypted in various ciphertext, this provides an additional security layer. But ElGamal algorithm has disadvantages: It increases ciphertext size twice than plaintext; Also as other asymmetric encryption algorithms ElGamal is slower and needs more computation time.

Normally, the ElGamal cryptosystem is used as an alternative for RSA cryptosystem and used in a hybrid cryptosystems. Because of fact, that ElGamal is slow during encrypting large amount of data, in hybrid cryptosystems ElGamal is used for the key encryption. Symmetric cipher are faster so they are used to encrypt plaintext [16, 17].

IV. RSA

RSA is very popular public-key encryption scheme. Among all asymmetric algorithms up to now, RSA algorithm is considered to be a secure and dependable. The RSA algorithm has a various lengths key. The security of RSA depends on

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

speed. RSA has variable length keys which change into 512-2048 bits [12]. According to different cryptanalysis, over the years RSA is considered as most reliable and secure algorithm among others.

Core component of RSA's security is how it is implemented and used [8]. As larger is key length, more secure is algorithm and it is harder to crack it through attack.

V. Comparison and Analyses of Some Symmetric and Asymmetric Ciphers

For experimental purposes was created JAVA program implementation on AES, Twofish, RSA ciphers and their hybrid models like Twofish+RSA and AES+RSA. This Java program implementation was used to research those algorithms. Using built Java code is possible to encrypt and decrypt data which is stored in the various sized text files. During the performance for each algorithm the program outcomes gives the information about energy efficiency (in terms of system memory usage) and the amount of time (displayed in Nanoseconds). Through experiments was used AES cipher, which is capable of handling 128-bit blocks.

To make researches on the Twofish algorithm for encryption is used Chilkat class Java library. This library for Twofish encryption uses variable length of key (128-bit, 192-bit, 256-bit). Library uses CBC (Cipher-Block Chaining) and ECB (Electronic Cookbook) modes. Presented Twofish uses 256-bits key length. During research used Twofish block size is 16 bytes, consequently encrypted output is always a multiple of 16.

The obtained program generates 2048 bits key pair for the RSA algorithm. This program can encrypt String type data with the public key and also decrypt this data with private keys.

As known, because RSA is a slow algorithm, it is rarely used to directly encrypt user data. Generally, RSA is used to encrypt shared keys, afterwards those keys are used for symmetric key cryptography. The advantage of this process is the speed of encryption decryption.

As for, Twofish algorithm, it is fast and easy to implement on different types of CPUs and hardware. Concerning technical usage Twofish has low requirements for applications. Twofish uses keys which are frequently changed, so it can work with little or no RAM and ROM available) [15].

Here is presented comparison of those algorithms according to their parameters spent during the decryption process. Program used during research has ability to work in different Unicode systems (For example in utf-8. It can decrypt and encrypt data in Georgian alphabet also). Table 1 table presents the results of research on a variety of sizes file.

Research shows, that throughout the encryption plaintext size increases proportionally with the time of encryption. Comparison results on algorithms AES-128 and Twofish regarding encryption time shows: For by encryption time criteria the Twofish algorithm is better and needs less time comparatively to AES algorithm; When AES needs approximately 3.178691038 times more encryption time than Twofish.

TABLE I. Twofish Encryption Decryption Process

Plaintext size (KB)	Twofish Encryption Time (nanoseconds)	Twofish Decryption Time (nanoseconds)	Twofish Encrypted file size (Bytes)	Used memory Bytes
32	1463712	1758853	65440	2241592
64	4140075	3235260	130848	2224680
128	5625297	4745870	261696	3271448
256	10012910	21730722	523392	5626680
512	20001135	18426348	1046752	10075240
1024	42106688	44281630	2096928	19001800
2048	81908320	129439314	4193856	33390848
4096	183173897	176583136	8387712	67292752

Also, Twofish and AES algorithms has difference values of the encrypted ciphertext size (bytes). Analysis of the obtained results makes obvious that Twofish increases encrypted file size approximately 6.2674 times more than AES (Fig.1.).

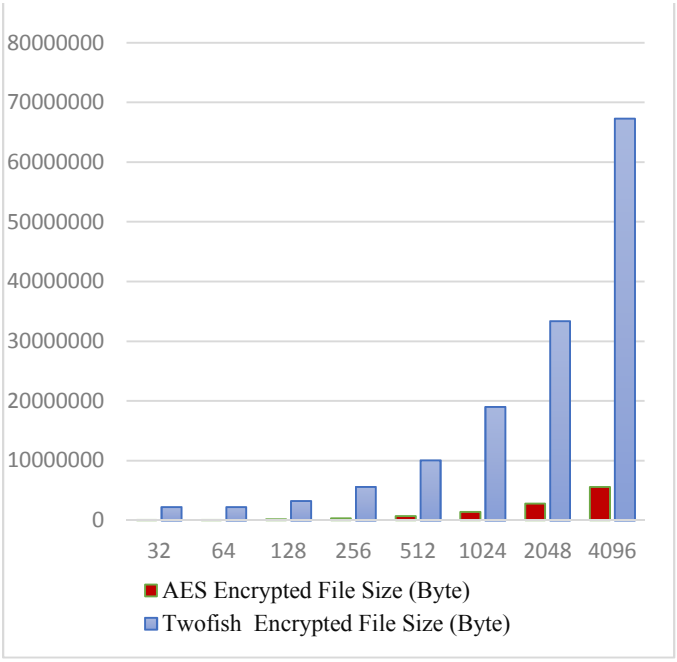


Fig 1. Comparison of AES and Twofish encrypted file sizes

The result of memory consumption analysis shows that the Twofish algorithm consumed 1.8468 times less memory than AES. Can be conclude that if Twofish works with small size data it needs fewer resources.

If we compare two public-key cryptosystems like RSA and ElGamal, we get the following results: when RSA is using the public key, ElGamal is slower than RSA. But ElGamal is faster when RSA decryption is using with a private key. Also, ElGamal has another advantage over RSA, file ElGamal-encrypted message is smaller than RSA-encrypted.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

VI. PROPOSED HYBRID SYSTEMS AND RESULTS OF EXPERIMENTAL RESEARCH

Hybrid encryption is called method which combines different numbers of encryption systems. Mostly, hybrid encryption merges asymmetric and symmetric encryption algorithms. It takes benefits from each used encryption system [17]. In general, using hybrid encryption ciphers remains public and private keys more protected, because of this hybrid algorithms are considered as less vulnerable type of encryption.

As we know, private-key algorithms are faster than public-key algorithms. Accordingly, proposed hybrid encryption systems security is also achieved with the slow public-key cryptosystem, using for key encryption, and fast private-key cryptosystem, which is used for plaintext encryption.

Fig.2. shows the general structural design of AES + RSA hybrid cryptosystem. In the beginning of the encryption process, the sender provides the plaintext (Message, which is presented in text file). Then, the system generates an AES key. This key must be encrypted with the RSA public-key algorithm. After this, the plaintext encryption process is done with AES cipher. Accordingly, the decryption process is reversed.

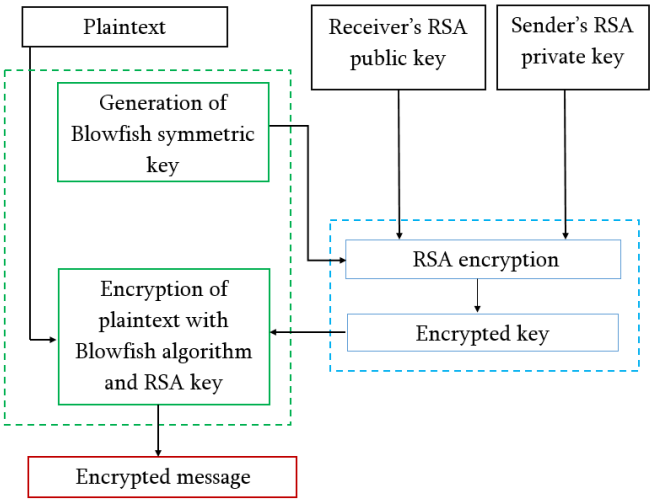


Fig 2. RSA + AES - the proposed hybrid system architecture

Similar method (see Fig.2) is used for the creation of Twofish+RSA hybrid algorithm. At first, the key will be encrypted with Twofish, and the message will be encrypted with the RSA 2048-bit key. At the end, the decryption process follows opposite order of the encryption process.

Conducted comparative analysis on AES, ElGamal and AES&ElGamal hybrid shows that new AES&ElGamal hybrid model needs less time than ElGamal. So, that means it is fast during encryption of different size text files. Proposed AES&ElGamal hybrid algorithm strength could be considered one of the competing with others [17].

From proposed algorithms were compared to detect which one was faster during encryption process: AES+RSA or AES. We have obtained the following results: that averagely AES cryptosystem is 0.9616 times faster, than AES+RSA hybrid model. (Fig.3.).

Table 2. Shows experimental results on different size files during encryption time.

TABLE II. ENCRYPTION PROCESS RESULTS OF PROPOSED HYBRID MODELS

Plaintext Size (KB)	AES + RSA Encryption Time (Nanoseconds)	Twofish +RSA Encryption Time (Nanoseconds)	AES+ElGamal Encryption Time (Nanoseconds)
32	372459621	4544527	22897995
64	477603056	8938838	26182346
128	501921935	15617402	32377061
256	529911194	30601857	47597261
512	570362261	49784217	68821938
1024	571026941	113360689	102385356
2048	588299138	223662075	139274046
4096	686185029	404410673	247789014
5120	816835306	544375155	301232405
6144	824170286	720005854	396602233

Proposed algorithms were also compared by memory consumption (in bytes). Comparison showed that Twofish+RSA hybrid model consumes less memory than others. After comparison presented algorithms can be sorted ascending as following: Twofish+RSA, AES+ElGamal, and AES+RSA.

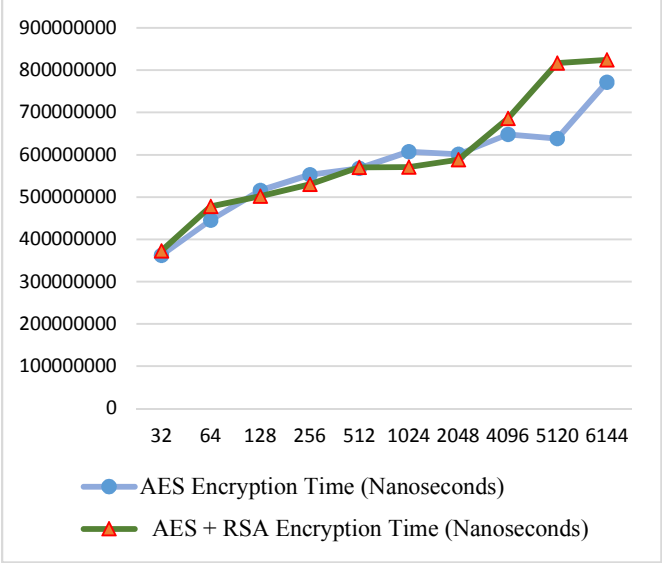


Fig 3. Comparison of AES and AES+RSA hybrid encryption time in nanoseconds

Making comparison between Twofish and Twofish+RSA algorithms in terms of encryption time shows that, the new hybrid model needs slightly more time (Twofish is 1.1775 times faster) than Twofish. (Fig.4.). During decryption process Twofish+RSA hybrid model is approximately 1.2772 times slower than Twofish. Twofish+RSA is slower because it contains public-key algorithm RSA which is slow in encryption.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

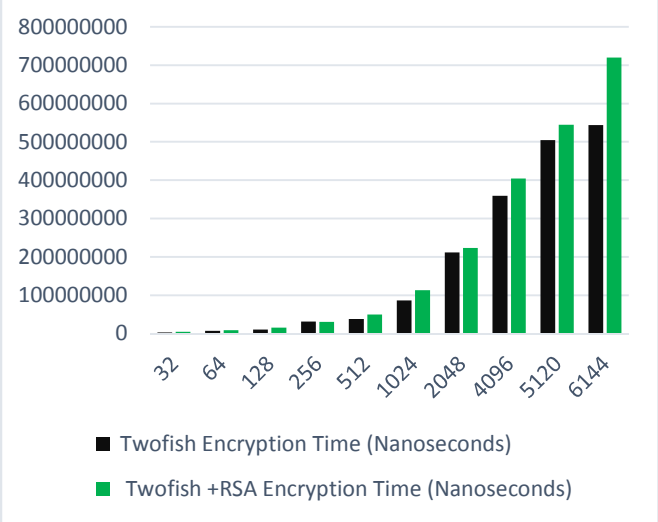


Fig 4. Comparison of AES and AES+RSA hybrid encryption time in nanoseconds

Taking into account the time and consumption of the technical resources, Twofish is the best one else the other described ones. After making comparison of Twofish, AES, Twofish&RSA, and AES&RSA hybrid algorithms according to the memory consumption results will show following: AES&RSA algorithm requires highest technical resources; Twofish&RSA is slightly forward than AES. (Fig. 5.)

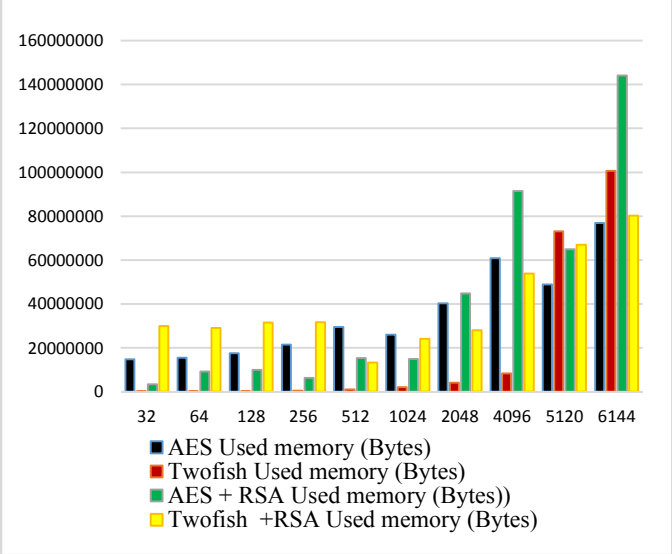


Fig 5. Comparison chart of memory usage during encryption process

Considering the encryption speed, the Twofish algorithm is the fastest among above analyzed systems. Although, Twofish + RSA algorithm during encryption needs less time and is faster, than AES+RSA. Among all those algorithms during encryption RSA is very slow (Fig.6). During the encryption process, AES+RSA needs 5.4932 times more time than Twofish+RSA. At the same time, the decryption process shows opposite results, and during the decryption process, AES+RSA is 2.35289 times faster than Twofish+RSA.

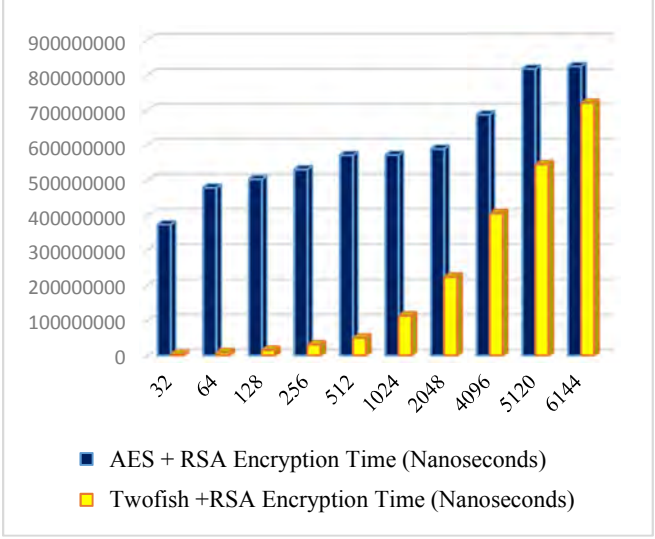


Fig 6. Comparison of TWOFISH+RSA and AES+RSA hybrid encryption time in nanoseconds

Proposed hybrid models were compared by encryption time (nanoseconds). They can be sorted acceding as following: AES+ElGamal, Twofish+RSA, AES+RSA. But while encrypting small files Twofish+RSA is faster, but with bigger files, AES+ElGamal is better (Fig.7.). During the research, AES+RSA gives medium numerical results but keeps a high-security level.

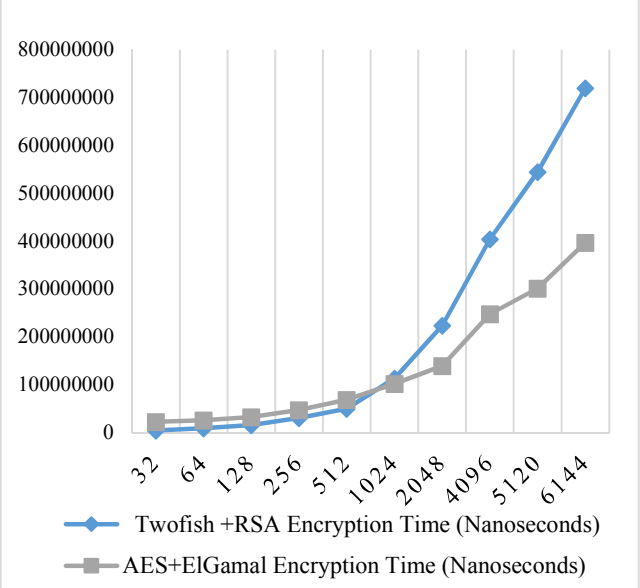


Fig 7. Comparison of TWOFISH+RSA and AES+ElGamal hybrid encryption time in nanoseconds

VII. CONCLUSION AND SCOPE OF FUTURE WORK

The paper presents comparative analyses AES, Twofish, RSA, and ElGamal cryptosystems. Based on those algorithms is created new hybrid cryptosystems Twofish+RSA, AES+RSA and AES+ElGamal. Memory consumption, encrypted file size, security level and encryption speed, those criteria were used to evaluate above proposed algorithms and hybrid models. After

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

research can be concluded that among the provided new hybrid models AES+RSA takes all benefits from used symmetric and asymmetric systems so it is significantly secure (Also AES was selected as a finalist for the Advanced Encryption Standard contest, NIST), but Twofish+RSA hybrid cryptosystem is faster.

For future work, proposed hybrid models can be analyzed by entropy index. With entropy research will be possible to evaluate resistance of each algorithm against different types of attacks, mostly against ciphertext frequency analysis.

REFERENCES

[1] Schneier B. Applied Cryptography: Protocols, Algorithms, and Source Code in C — John Wiley & Sons, 1996.

[2] Kelsey J., Schneier B., Wagner D. (1996). Key-schedule cryptanalysis of IDEA, G-DES, GOST, SAFER, and Triple-DES

[3] Meyers R. K., Desoky A. H. An Implementation of the Blowfish Cryptosystem // Signal Processing and Information Technology, 2008. ISSPIT 2008. IEEE International

[4] Symposium on — Institute of Electrical and Electronics Engineers, 2008

[5] R. L. Rivest, A. Shamir and L. Adleman. “A method for obtaining

[6] Digital signatures and public-key cryptosystems,” Comm. ACM, 21, pp. 120-126, 1978.

[7] Schneier B. Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish) // Fast Software Encryption: Cambridge Security Workshop Cambridge, U. K., December 9–11,1993 Proceedings / R. J. Anderson — Berlin: Springer Berlin Heidelberg, 1994

[8] Шнайер Б. (2002). Прикладная криптография. Протоколы, алгоритмы, исходные тексты на языке Си — М.: Триумф

[9] Shiho Moriai; Yiqun Lisa Yin (2000). "Cryptanalysis of Twofish (II)"

[10] Bruce Schneier; Doug Whiting (2000-04-07). "A Performance Comparison of the Five AES Finalists"

[11] Schneier, Bruce (2005-11-23). "Twofish Cryptanalysis Rumors". Schneier on Security blog.

[12] «Announcing Request for Candidate Algorithm Nominations for the Advanced Encryption Standard (AES)» (англ.). Department of Commerce — National Institute of Standards and Technology — Federal Register: September 12, 1997

[13] Niels Ferguson (1999-10-05). "Impossible differentials in Twofish".

[14] R. L. Rivest, A. Shamir and L. Adleman. “A method for obtaining digital signatures and public-key cryptosystems,” Comm. ACM, 21, pp. 120-126, 1978.

[15] The Twofish Encryption Algorithm, B. Schneier, Dr. Dobb's Journal, December 1998.

[16] Taher ElGamal (1985). «A Public-Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms

[17] 2018 IEEE 5th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC) - "Hybrid encryption model of AES and ElGamal cryptosystems for flight control systems", Maksim Iavich, Elza Jintcharadze, Sergiy Gnatyuk, Yuliia Polishchuk and Roman Odarchenko

[18] Johhanes A. Buhman, Introduction to Cryptography, Second Edition, 2000

[19] Alfred J. Menezes, Paul C. van Oorschot, Scott A. Vanston, Handbook of Applied Cryptography, Massachusetts Institute of Technology, June 1996

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

The Integrated Approach to Automation and Digitalization of the Transport Processes in the Industrial Enterprises

Alexey G. Lekarev,
Chairman of the Board of Directors of Vega LLC
St. Petersburg, Russia
lagsbor@mail.ru

Maxim G. Ammosov,
General Director of Vega LLC
St. Petersburg, Russia
m.ammosov@tshss.ru

Dmitry V. Efanov,
D. Sc., Associate Professor, First Deputy General Director – Chief Engineer of Vega LLC
St. Petersburg, Russia
TrES-4b@yandex.ru

German V. Osadchy,
Technical Director of Vega LLC
St. Petersburg, Russia
osgerman@mail.ru

Natalia A. Goncharova,
Engineer of Vega LLC
St. Petersburg, Russia
nataliegoncharova@list.ru

Abstract—The paper proposes the architecture of the integrated control system for industrial enterprises, covering the transport and production components. The role of technical means of monitoring the objects parameters of implementation of production and transport processes at the enterprise is noted. The subsystems for monitoring the technical condition and functioning parameters of infrastructure facilities and the rolling stock are the central links in the structure. The presence of such subsystems makes it possible to synthesize at the central post level the models of the functioning of each component of an industrial enterprise and to develop recommendations for optimizing the industrial production. This paper describes, for example, the ways of implementing the subsystem for monitoring the location of the rolling stock in the railway transport systems of the industrial enterprises. The use of the structure presented in the paper as applied to each particular enterprise requires the technical examination and the detailing to particular participants in the technological processes of the enterprise. The presence of an integrated control system makes it possible to solve the main problems of the enterprise, associated primarily with the inability to implement (or low speed of implementation) of the transportation process at peak loads and with a low coefficient of machine use at the enterprise. The implementation of the integrated control systems is an important stage in optimizing the activities of industrial enterprises.

Keywords—*transport system of industrial enterprises; railway traffic optimization; integrated control system; monitoring and forecasting of the traffic, parameters, and the condition.*

I. Introduction

The development of science and technology in the first quarter of the XXI century makes it possible to solve the most complex problems, such as fully automatic control of transport units and forecasting changes in the operation. Modern aircraft are widely used for automatic piloting [1], autopiloted vehicles are being developed and tested [2-4], etc. Many transport applications use artificial intelligence technologies to analyze the technical condition of infrastructure objects and the rolling stock in order to predict changes in the technical condition and, as a result, assess the residual life of the functioning [5 – 8]. Humanity has long learned to get digital pictures for a variety of objects, both from everyday life and from the field of the responsible technological processes control.

The most important stage in the development of engineering and technology is the use of integrated, "end-to-end" technical solutions for control large, separated industrial and transport enterprises. Most owners of infrastructure and the rolling stock do not fully use such systems now. The use of automation tools is limited only to the production processes of the enterprise. The lack of communication between the transport processes control systems and the production processes control systems of enterprises does not make it possible to optimize the operation of the entire complex. First of all, this does not make it possible to increase the coefficient of machine use and to provide a non-stop transport process at peak loads on the transport system. In other words, the owner has a hidden resource for increasing the capacity of the traffic and the implementation of processes – the use of technical means for monitoring and optimizing processes in the enterprise.

The purpose of this paper is to introduce the reader to the architecture of the integrated control system of transport systems of industrial enterprises.

II. The Scheme of Interaction of the Enterprise Objects

Modern industrial enterprises are equipped with various technical means of automation of production and transport processes [9]. The choice of methods and technical means of automation is determined on the basis of the company's own policy and priorities, taking into account the need for the best implementation of the main production processes. In addition to technical means of industrial automation, many industrial enterprises have their own developed transport systems, including roads and railways, as well as (in some cases), and water transport terminals.

978-1-7281-9899-6/20/$31.00 ©2020

An analysis of the specifics of a large number of industrial enterprises showed the following important feature of their functioning: almost all enterprises have automated control systems for transportation that are not directly automatically linked to the control systems of production enterprises. In other words, the transport system is being "picked up" by the managerial staff of the means of automation of the enterprise's production processes. The operation of the main means of sales and production is supplemented by the operation of the transport system, and requests from the "production workshop" are requests for the operation of its machines. In some cases, the lack of technical means for monitoring the state and operation parameters of transport system facilities and optimizing its operation leads to two major problems: the inability of implementing (or low speed of implementation) of the transportation process at peak loads and a low coefficient of machine use at the enterprise.

The company may have its own, rented, and visiting automobile vehicles, mostly not equipped with technical means for monitoring the condition and the movement. Exceptions may be objects of special equipment. If the enterprise has a railway complex, it usually combines zones of automated traffic control with the peripheral centralization posts and manual control zones. The first ones are equipped with an alarm system and auxiliary remote control of floor automation objects (arrows and traffic lights) from the centralization posts to transmit information to train drivers, and the second ones imply the traffic only with the participation of technical personnel of the infrastructure complex and train drivers. It should be noted that the technical means of the transportation process automatization on the railways of industrial enterprises correspond to the highest level of safety integrity (as well as on mainline railways) [10]. It should be noted that any technical means of signaling, centralization and blocking in industrial enterprises (as well as on mainline railways) are point-based solutions for remote control of floor objects, but not means of automation of operational planning, control and dispatching of transport processes. The roles of the latter ones are performed by special technical personnel, who carry out their work on coordinating the actions of station attendants and train drivers, keep records of executed train and freight operations, document support of transport processes, etc.

A huge leap in the field of information and computer technologies makes it possible now to talk about complex technical solutions for the automation of production and transport processes of industrial enterprises. The Fig. 1 shows a structural diagram of the interaction of subsystems of such a complex control system, which is universal and covers both the transport complex of any enterprise (both road and water transport, and the railway), and the complex of technical means for automating the main production processes of the enterprise.

In our opinion, the key disadvantages of modern transport systems of any industrial enterprise are the lack of developed technical means for monitoring the processes and the state of its participants, as well as the lack of means for optimizing the enterprise processes, taking into account the real situation at the production and transport facilities.

All equipment of an integrated control system can be classified into the equipment of the central posts of the process dispatching, the equipment of the infrastructure complex, as well as the equipment of the rolling stock. All equipment is in direct control or in information interaction. Total automation of all processes is not intended, but it is intended to minimize manual labor and reduce the human operator to the role of an observer in many components of transport systems. To a certain extent, this also applies to the operation with the rolling stock, which can be implemented with direct control of a person from the board (from the driver's cab, from the driver's seat in a car or in a special equipment object), using the distance means of the control from the panel and, in the longer term, automatic traffic control by the central control system. It is assumed to minimize the impact on the rolling stock.

The main means of equipment of the rolling stock are technical means of the control and tools for monitoring the status and processes, as well as movement parameters. The technical means of the control supplied by the manufacturer of the rolling stock include only the means of the direct influence on the control aggregates of the transport unit itself. To implement an integrated control system of the enterprise, it is necessary to equip each traction (self-propelled) unit with technical means for monitoring of the infrastructure, its own condition, the processes occurring in it, as well as the movement parameters. Non-self-propelled units are equipped with simpler monitoring tools to control their movement by traction units. To implement such monitoring systems, specialized sensors for taking physical parameters and linking objects to an electronic map of the enterprise are installed (if such a map is not prohibited by the conditions of the enterprise). The sensors of non-self-propelled units are completely autonomous (include high-capacity batteries, sources of alternative energy supply, remote recharging, etc.) and transmit information through an expandable radio channel or mobile and satellite navigation. It is advisable to transfer data on non-self-propelled units to the technical means of on-board automation of self-propelled units that have replenished sources of energy supply, and it is also much easier to communicate with the equipment of the central posts. The presence of sensors for the status of the rolling stock and control of their location makes it possible to construct the models of their movement in the software of the central posts, as well as to compare the executed movement with the planned one, by correcting the modes of control, taking into account the real transport and industrial situation.

Stationary monitoring tools are also used to improve the positioning accuracy of the rolling stock. It also assumes the implementation of the function of monitoring the technical condition of infrastructure facilities. It is also necessary to monitor the main production processes and transfer data about them to optimize transport processes. Stationary monitoring tools can be implemented with cable data transmission and cable power, which simplifies their operation.

The technical solutions with alternative energy sources are possible, including the use of "green technologies" [11]. The data from the stationary monitoring tools, directly or through the equipment of peripheral posts of the control centralization, is transferred to the technical means of controlling the production processes of the enterprise, as well as to the technical means of dispatching transport processes. From there, the data is transmitted to the means of optimizing the enterprise's processes and, in particular, the transportation process.

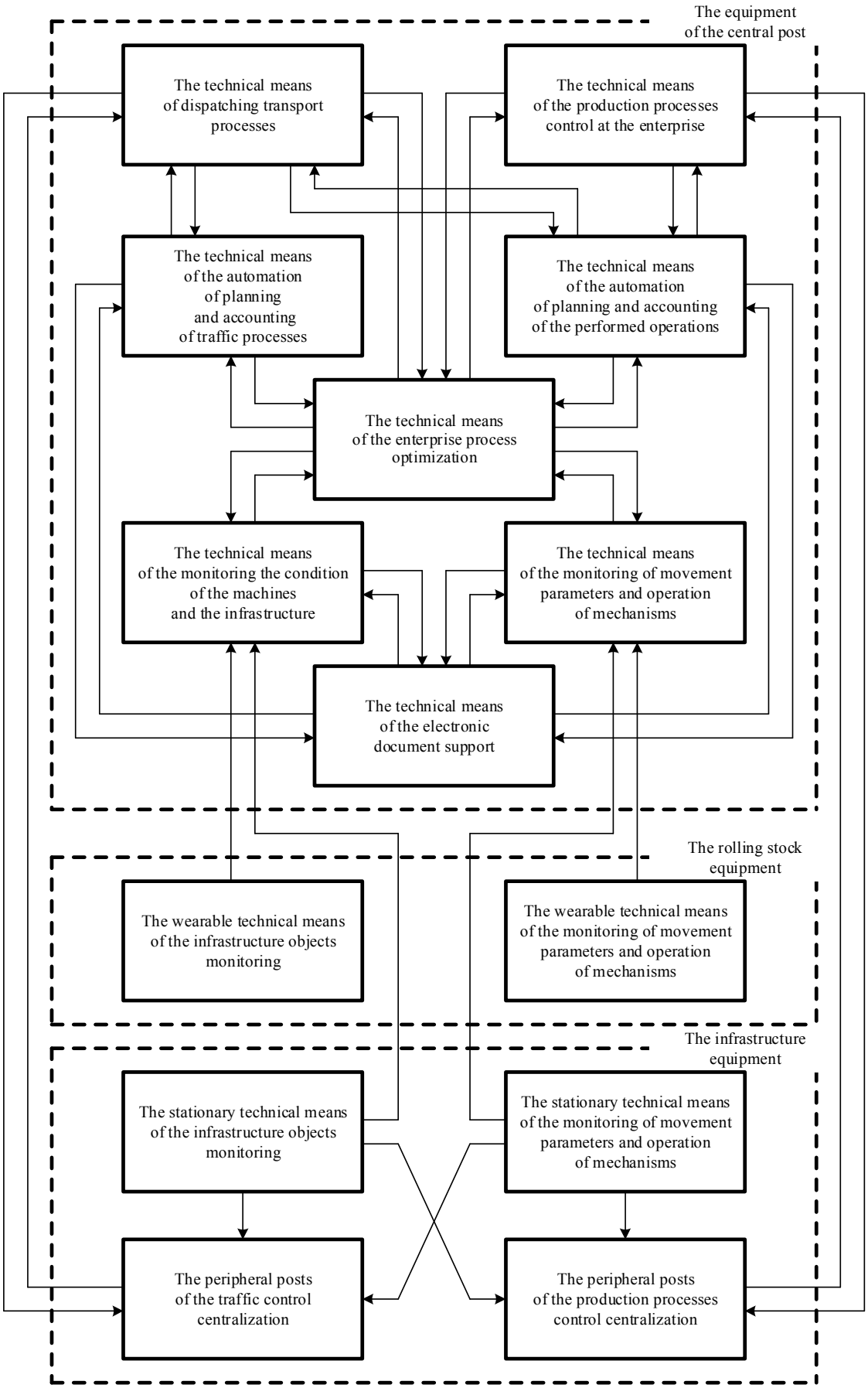


Fig. 1. The organizational diagram of the interaction between the components of the control and monitoring systems at industrial enterprises.

For each object covered by monitoring sensors, an intellectual analysis of the technical condition, assessment of the actual load and the cost of the operation, as well as the predicted residual life, taking into account the real load, etc. is possible. This also applies to infrastructure facilities and the rolling stock.

III. The Approach to the Automation of the Motion Models Construction in the Railway Transport System

We give an example of the implementation of the level of control of the location of non-self-propelled units from the field of railway transport [12]. Non-self-propelled units (cars) are equipped with autonomous sensors (tags). This tag is attached to the car when it hits the enterprise, it makes it possible to track the location of the car when it runs inside the railway transport system. When a car is sent outside the enterprise, the tag is dismantled. A self-propelled unit (locomotive) is also equipped with a tag, functioning from the technical means of on-board automation and power supply and carrying out a survey of car tags (Fig. 2).

Car tags are devices that are fixed on the rolling stock by simple attachment (for example, on clamps or staples) and are completely autonomous. High-power batteries are used for their energy supply, and data transfer is based on the technology of the Industrial Internet of Things (IIoT) [13 – 15]. Car tags always work “under load”, and “wake up” when exposed to an active locomotive tag with confirmation of the binding of cars or a group of cars, and then “fall asleep” again. At the same time, the specifics of the enterprise are taken into account in the operation. For example, in the mining industry, dump trucks are often connected to the so-called car "turntables", including dozens of cars, and are rarely uncoupled in the presence of defects. In this case, it is possible to reduce the number of car tags and use them only for the extreme cars that are located in close proximity to the traction unit. These numbers are manually assigned in the technological windows of the automated workplaces of freight work to the numbers of the cars included in the “turntable” in a strict sequence of their positions relative to each other. If irregularities are revealed in the work with cars, the cargo operator will correct the numbers according to the actual situation.

To determine the location of the rolling stock, the territory of the enterprise is covered by base stations, which are connected to the receiving and transmitting devices of locomotives. To improve the accuracy of the rolling stock positioning, it is linked to stationary train traffic control systems [16]. At the same time, it is linked to existing systems, and in new construction, it is possible to retrofit railway tracks with sensors for monitoring the position of the rolling stock (for example, axis counting sensors, including fully autonomous ones) [17]. Data from existing automation systems is obtained by docking according to a specialized protocol, which requires the installation of an additional data output board in the hardware of microprocessor centralizations, and when using relay centralizations, it requires the installation of controllers for receiving discrete data from remote control panels and control devices. Well-known industrial automation controllers can be used as such controllers. An alternative is the modification of existing relay-type systems into relay-processor-type systems, where in fact the control apparatus and part of the circuit solutions for transmitting commands to the executive level are implemented using computer technology. The use of purely microprocessor-based centralization seems redundant, because the resource of relay systems is much (by an order of magnitude) higher.

The locomotive tag is a component object of the onboard control system. When performing a docking operation with a particular car or a group of cars, the active tag generates a wake-up signal for the car tags, which transmit their identifiers and go into the sleep mode again. The on-board system remembers this group of cars (or particular cars), passing them to the car and train models, identifying them as a group of cars “belonging” to the on-board tag. During the uncoupling operation, the locomotive tag generates a wake-up signal for the cars tags. These operations are recorded in the car and train models in the software of an integrated control system.

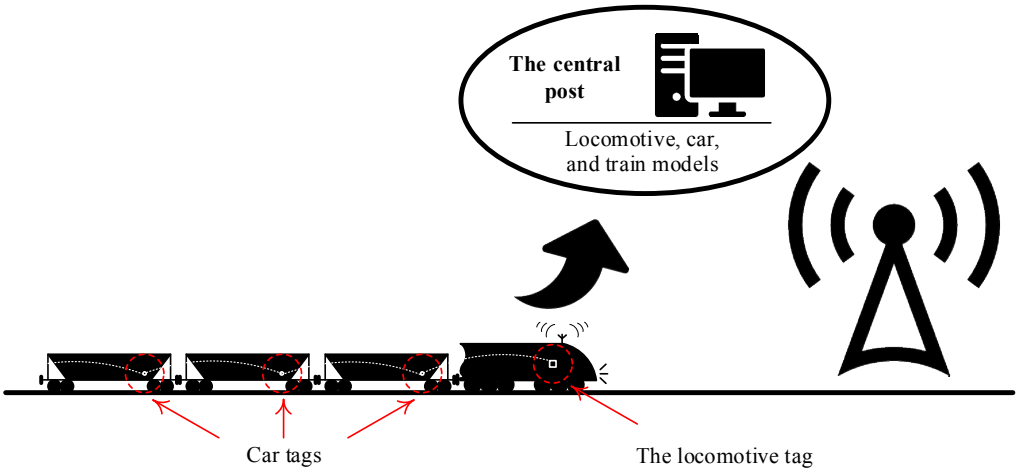


Fig. 2. The organizational structure of the control system.

The presence of car and locomotive tags makes it possible to "revive" the cars, train and locomotive models in the software of an integrated control system. All this is implemented, including linking to electronic cards. In addition, it is possible to automatically maintain a schedule of executed traffic without installing separate technical solutions according to the type of solutions on the main railways, as well as automatic comparison of planned and executed schedules with a possible automatic prompt to the dispatcher of correction options for traffic processes.

At various enterprises, locomotives can be rented and own, therefore, the on-board automation complex should be modular and removable and installed in the cabin of a particular locomotive for the duration of operation. At this time, such a module is not remote and is fully identified with the locomotive on which it is installed. Modules also include the base stations. For stabile operation, a locomotive antenna is also used, which provides a high-quality signal with stationary equipment and tags.

Thus, the locomotive tags must be configured to the operation with a specific base station located on it. To check the coordinates of uncoupled cars (or whole "turntables"), tags must have a connection mode every certain period (for example, every 10 minutes). The locomotive base station should be able to work both with stationary base stations and with local tags located on cars.

The specified set of solutions makes it possible to link at the level of the central post all three models of the rolling stock: locomotive, car and train models. Information about it is basic for solving other tasks.

IV. Conclusion

The structural scheme of the integrated control system of the industrial enterprise proposed in this article is universal and covers both the automation system of the main production processes of the enterprise and the transport system without the traditional separation of the water, road or rail system. The focus of the system is on quickly obtaining the objective data about the technical condition, operation parameters and location of participants and the main technological process and transportation process. This makes it possible to implement a digital picture of the enterprise and develop solutions to optimize processes, increasing both production efficiency and product transportation efficiency.

The main components of an integrated control system are peripheral tools for monitoring the parameters of functioning and technical condition of the enterprise's infrastructure and the rolling stock. Due to their presence in the software of technical means of control at the central level, all the data necessary for solving the problems of rational production and transportation control is generated.

For the real enterprise, when detailing the proposed structure, a technical examination and the formation of a conceptual plan for the development of the enterprise are required, taking into account existing control systems and opportunities to increase the production efficiency.

References

[1] L. Deka, and M. Chowdhury "Transportation Cyber-Physical Systems", Elsevier, Amsterdam, Netherlands, 2019, 348 p.

[2] P. Sharma, H. Liu, H. Wang, and S. Zhang "Securing Wireless Communications of Connected Vehicles with Artificial Intelligence", IEEE International Symposium on Technologies for Homeland Security (HST), 25-26 April 2017, Waltham, MA, USA, doi: 10.1109/THS.2017.7943477.

[3] M. Dikmen, and C. Burns "Trust in Autonomous Vehicles: The Case of Tesla Autopilot and Summon", IEEE International Conference on Systems, Man, and Cybernetics (SMC), 5-8 October 2017, Banff, AB, Canada, pp. 1093-1098, doi: 10.1109/SMC.2017.8122757.

[4] B. Brown "The Social Life of Autonomous Cars", Computer, 2017, vol. 50, issue 2, pp. 10.1109/MC.2017.59.

[5] T. Asada "Novel Condition Monitoring Techniques Applied to Improve the Dependability of Railway Point Machines", University of Birmingham, UK, Ph. D. thesis, May 2013, 149 p.

[6] W. Jin, Z. Shi, D. Siegel, P. Dersin, C. Douziech, M. Pugnaloni, P. La Cascia, and J. Lee "Development and Evaluation of Health Monitoring Techniques for Railway Point Machines", 2015 IEEE Conference on Prognostics and Health Management (PHM), 22-25 June 2015, Austin, TX, USA, DOI: 10.1109/ICPHM.2015.7245016.

[7] T. Böhm "Remaining Useful Life Prediction for Railway Switch Engines Using Artificial Neural Networks and Support Vector Machines", International Journal of Prognostics and Health Management 8(Special Issue on Railways & Mass Transportation), December 2017, pp. 1-15.

[8] L. Heidmann "Smart Point Machines: Paving the Way for Predictive Maintenance", Signal+Draht, 2018 (110), no. 9, pp. 70-75.

[9] I.I. Barankova, G.I. Lukianov, and U.V. Mikhailova "Company Railway Transport Control Automation", 2017 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 16-19 May 2017, St. Petersburg, Russia, DOI: 10.1109/ICIEAM.2017.8076138.

[10] D.J. Smith, and K.G.L. Simpson "Functional Safety: A Straightforward Guide to IEC 61508 and Related Standards", Butterworth-Heinemann; 1st edition (June 26, 2001), 208 p.

[11] D.W. Efanow, and G.W. Osadtschiy "Energy Efficiency Categories for Safety Installations", Signal+Draht, 2020 (112), no. 4, pp. 36-42.

[12] D.V. Efanov "The Principles of Automation of Traffic Control Processes on the Railways of Industrial Enterprises" (in Russ.), Transport of the Russian Federation, 2019, issue 6, pp. 27-33.

[13] I. Barankova, U. Mikhailova, and G. Lukianov "Automated Control System of a Factory Railway Transport Based on ZigBee", 2016 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), 19-20 May 2016, Chelyabinsk, Russia, DOI: 10.1109/ICIEAM.2016.7910923.

[14] 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 15th IEEE East-West Design & Test Symposium (EWDTS`2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 231-236, doi: 10.1109/EWDTS.2017.8110152.

[15] V. Hahanov "Cyber Physical Computing for IoT-driven Services. – New York, Springer International Publishing AG, 2018, 279 p.

[16] G. Theeg, and S. Vlasenko "Railway Signalling & Interlocking: 3ed Edition", Germany, Leverkusen PMC Media House GmbH, 2020, 552 p.

[17] J. Exner, and P. Pohl "The Potential and the Development Focus for Distributed Fibre Optic Sensing in the Rail Sector", Signal + Draht, 2019, (111), no. 1+2, pp. 27-30.

A recommender subsystem construction for calculating the probability of a violation by a locomotive driver using machine-learning algorithms

Valentina Sidorenko
Professor of Department "Control and Information Security"
Russian University of Transport (MIIT) and Department Business Process Modeling and Optimization Higher School of Economics
Moscow, Russia
valenfalk@mail.ru

Maksim Kulagin
Postgraduate of Department "Control and Information Security"
Russian University of Transport (MIIT)
Moscow, Russia
maksimkulagin06@yandex.ru

***Abstract* — This article describes the issues of analysis and assessment of the human factor for predicting the violation committed by the locomotive driver when driving the electric rolling stock. An intelligent system overview for assessing the likelihood of a violation by a locomotive driver is given. Such a system can generate recommendations depending on previously committed violations. One of the tasks is to reduce the risk of locomotive safety devices malfunctions, which are part of the locomotive electrical equipment. The solution to the problem of predicting the occurrence of possible violations is solved using tools and machine learning algorithms. A model has been built that generates recommendations for the driver based on information about previously committed violations and several static characteristics of the locomotive driver.**

Keywords — electric rolling stock, machine learning, recommender algorithms, neural network, optimization.

I. Introduction

Nowadays, machine learning methods are becoming increasingly popular in computer science. They allow solving the problems of analyzing, predicting, detecting and recognizing.

Self-driving car systems are actively developing and popularizing all over the world. However, in large transport companies, transportation is still performed by people. Professional drivers are needed to ensure high quality and safety of transportation. Many companies train their employees on their own and then control their work, for example, at the Russian Railways company.

II. Background

In this study, machine learning methods and algorithms were used. The analysis of machine learning methods applied to the field of railway transport and the human factor was carried out.

Many scientists are working on solving problems associated with diagnosing hardware of locomotives on railways. There are solutions for predicting and managing the state of the diagnostic object based on the estimated time to failure [1]. An overview of forecasting methods is presented in [2], the methods are classified depending on the required computing resources and the amount of historical data. A comparison of different data-driven approaches such as principal component analysis (PCA) and partial least squares (PLS) is given in [3].

Russian Railways is actively using systems based on deep learning algorithms. The main tasks of the scientific and engineering community are the detection and recognition of objects on the railway in images and videos [4, 5, 6].

There are several examples of using machine learning techniques to detect anomalies and find faulty electrical equipment. In [7], the authors use methods and approaches to regression analysis from the field of digital signal processing for troubleshooting and deviations in the operation of electrical appliances. In [8], the authors present anomaly detection methods based on a combination of nonparametric statistical testing and machine learning methods and demonstrate the effectiveness of an anomaly detection strategy using real operational data of locomotives.

To solve problems related to the human factor, machine learning methods are widely used both on the railways and in industry. There is an adaptive management system for the railway infrastructure maintenance (URRAN PROJECT) of Russian Railways [9, 10, 11], but there is no objective system for assessing the locomotive driver performance. The creation of a system for the automatic formation of a comprehensive assessment of the locomotive driver performance is an urgent scientific and technical task at Russian Railways. Analysis of human activity by machine learning is also used in medicine and banking industry. For example, in medicine, a diagnostic system using neural networks is becoming popular, which relies on human medical indicators and mathematical apparatus [12, 13]. In the banking sector, the problem of credit scoring and credit analysis of a bank is investigated and solved [14].

Detailed analysis and description of the constructing model's principles for the formation of various objects ratings of research in the field decision theory are presented in [15, 16].

There is a methodology for assessing and methods for reducing human error [17], empirically confirmed. In practice, this method is used to quantify the likelihood of error during the execution of a production task using weighting factors.

Methodology for assessing the reliability of railway processes [18] for a specific approach to quantifying human errors. The main aim of the method is to improve accounting for human reliability and provide a simple tool for quantifying human error in the railway industry.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

III. TASK

The diagram shows the factors affecting the safety of operation, the locomotive technical condition, compliance with the operating and maintenance rules (Fig. 1). Among these factors, the human factor is determining factor [19, 20, 21, 22]

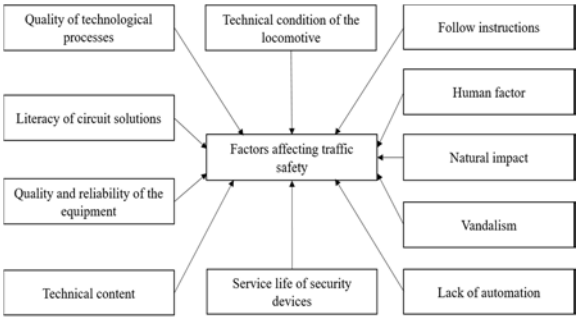


Fig. 1. Factors affecting security

The introduction of an intelligent system for a comprehensive assessment of the likelihood of a violation by a locomotive driver and the formation of recommendations, depending on previously committed violations on the railway, allows reducing the influence of the human factor on the safety of operation. Currently, there are no such systems on Russian railways. This determines the practical significance of this study. Methods of expert assessments [17, 18, 23, 24], Markov models [25], cognitive models [26] are used to solve such problems in the world. Locomotive drivers commit quite a few violations affecting both the condition of the rolling stock and safety of operation (Fig. 2). About 35% of driver violations in 2020 are related to brake control (Fig. 3).

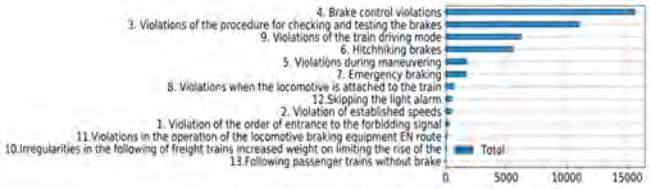


Fig. 2. Distribution of violations committed by the driver for 2020 by groups

Fig. 3. Classification of violations committed by drivers by the group «The violation of the brake control»

It should be noted that the group "Violation of the procedure for approaching a restrictive signal" includes violations that have a very high risk in terms of the possibility of locomotive collision or running off the rails (Fig. 4). Violations of this group are considered without the locomotive driver's error only in 39.6% of cases.

Fig. 4. Classification of violations committed by drivers by group «Violation of the order of access to the prohibiting signal»

Research and creation of a unified and objective methodology for a comprehensive assessment of the likelihood of a locomotive driver's violation and the formation of recommendations depending on previously committed violations is aimed at improving the safety of operation on railway transport. The purpose of this study is to develop software (mathematical models, methods and algorithms) to determine the likelihood of a driver committing a violation and form a list of measures to reduce it.

IV. APPROACH

The design process of an intelligent system for a comprehensive assessment of the probability of a locomotive driver committing a violation and making recommendations, depending on previous violations, can be divided into several stages (Fig. 5):

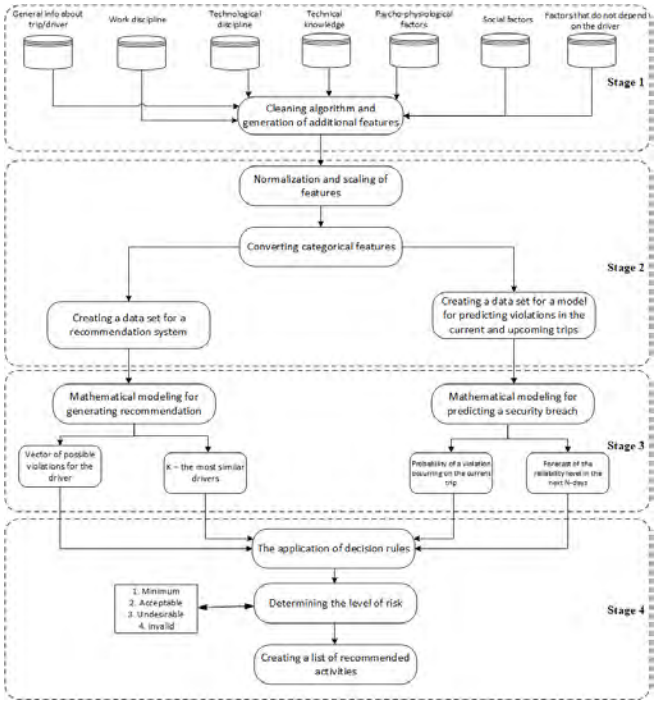


Fig. 5. Block diagram of an intelligent system

Stage 1. Determination and unloading of the list of necessary indicators from the automated control system, which characterize the work of the foot-plate staff (crew), adjustment of data, elaboration.

Stage 2. Qualitative data processing - the transformation of categorical features, scaling and normalization of numerical features, search for linear and nonlinear-dependent features.

Stage 3. Design of a complex algorithm for calculating the committing a violation probability, calculating the reliability level, building a rating of foot-plate staff (crew) and generating recommendations for each locomotive driver.

Stage 4. Building a decision function - combining the various algorithms results, dividing locomotive drivers into reliability groups and automating the formation of a list of measures to reduce the risk level.

In the course of studying the data on violations committed by locomotive drivers, it was revealed that the appearance of violations entails the appearance of similar ones. Graphical illustrations demonstrate this (Figure 6).

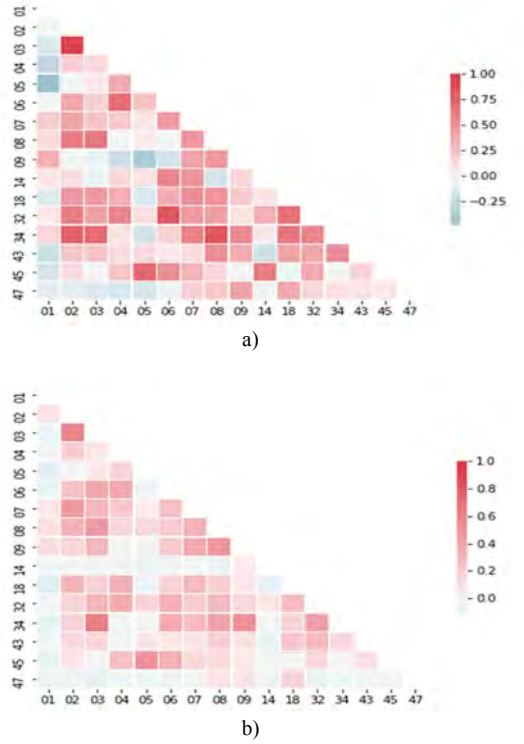


Fig. 6. Correlation between groups of violations on roads (a) and depots (b)

About 40 types of violations have a correlation coefficient of 0.5 to 1. This is approximately 12% of the violations types of the total number. Violations from the group *"34 - Untimely activation of locomotive safety devices before departure"* have a high correlation (0.6 - 0.7) from the group *"2 - Violation of the order of approach to the restrictive signal"* and *"3 - Violation of the set speeds"*. Violations from the group "*43 - Driver violations during the operation of locomotive safety devices*" have a high correlation (0.4 - 0.5) with the groups *"18 - automatic train stop"* and *"32 - Short-term disconnection of the main locomotive safety devices along the route"*. The correlation between violations *"09.3 - Lack of checking the brake line tightness"* and *"10.7 - Lack of blowing the brake line when accepting the locomotive"* is 0.66.

Proceeding from the fact that the occurrence of violations leads to the appearance of similar (Fig. 6) and dependencies between the locomotive driver's indicators and the violations committed [21], a model is built based on an intelligent system. The model building process can be divided into three stages.

1. The first step was to build a rating matrix *R*, where the row is the locomotive driver and the column is the violation number. The values are the number of points scored by the driver for violations over a certain period.

2. The second step is the decomposition of the matrix ratings into a matrix smaller rank Eq. (1):

$$R \approx UV^T, \tag{1}$$

where U – matrix of locomotive drivers, V – matrix of violations displayed the correlation between both drivers and violations at the same time (Fig. 7).

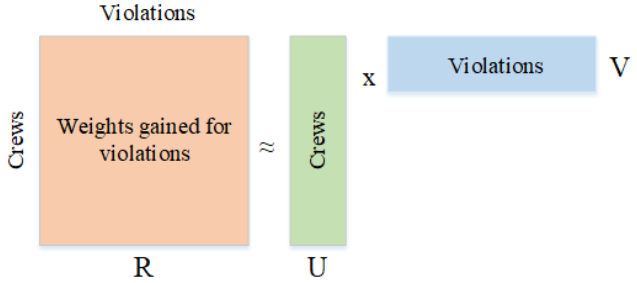


Fig. 7. Splitting a matrix into two with a lower rank

The decomposition is performed using an algorithm ALS (Alternating Least Squares). ALS performs an iterative optimization process. For each iteration, it tries to approach the factorized representation of our source data [27]. The optimization problem for this decomposition is to minimize the loss function *J Eq.* (2):

$$Minimize\ J = \frac{1}{2}||R - UV^T||^2. \tag{2}$$

3. At the third stage, the architecture is formed, and the neural network is trained to predict the probability of committing each of the possible violations. The locomotive driver's feature and the results of the decomposition of the interaction matrix into *U* and *V* matrices are used as attributes

The figure below (Fig. 8) shows a neural network in which the input layer is a feature vector formed from 3 different data sets (*U*, *V*, and other driver feature). The hidden layers of the neural network use the *ReLU Eq.* (3), activation function and the output layer use *Softmax Eq.* (4).

$$ReLU_j^l = max\left(0, a_j^{[l-1]}\right), \tag{3}$$

$$Softmax_j = \frac{e^{a_j^{[L-1]}}}{\Sigma_{i=1}^{n} e^{a_i^{[L-1]}}}, \tag{4}$$

where l – index of the hidden layer; L – index of the output layer, number of layers; i – index of neuron in the layer; n – number of neurons in the layer; $a_j^{[l-1]}$ – output activation function j^{th} neuron $l-1$ layer.

The neural network hyperparameters used for selecting parameters by the gradient descent method are as follows:

– learning rate – 0.1;
– number of hidden layers – 3;
– number of neurons in each layer – [64, 32, 16];
– additional feature – 11;
– epochs steps – 500.

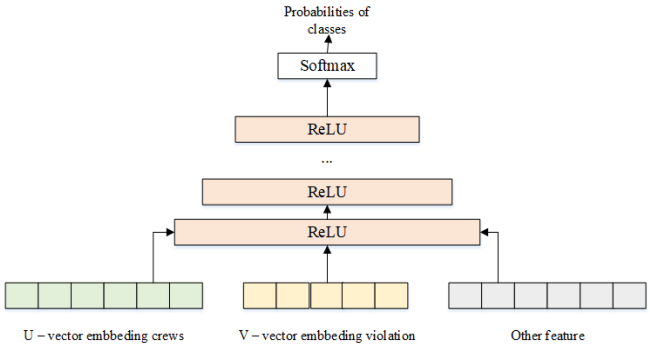


Fig. 8. The neural network of the recommendation subsystem based on the results of the decomposition of the interaction matrix and static signs of the driver

V. RESULTS

The neural network was tested on a test sample of 5-10 thousand locomotive drivers. Since the accuracy *Eq.* (5) for predicting the violation class is low [27, 28], modified metrics were used to assess the quality, namely accuracy of getting violations in the top 10 and 20 violations in the predicted vector $Accuracy@topK$ *Eq.* (6):

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}[c(x_i) = y_i], \quad (5)$$

$$Accuracy@topK = \frac{1}{N}\sum_{i=1}^{N}[y_i \in c_{topK}(x_i)], \quad (6)$$

where N – number of examples in the sample; y_i – fact class; $c(x_i)$ – predicted class by the model for the object x_i ; $c_{topK}(x_i)$ – set of $topK$ the most likely classes predicted for the object x_i; $topK$ – hyperparameter that shows whether the actual class occurs among the K most likely classes predicted by the model.

The results of the model are presented below (TABLE I.)

TABLE I. RESULTS OF THE MODEL FOR A TEST SAMPLE

Accuracy	Accuracy@top10	Accuracy@top20	Loss
0,112	0,439	0,617	4,241

After building the model, you can identify several violations that were predicted most often (TABLE II.).

The quality of the forecast is low because the forecast is made for each violation separately. This article refers to a multicriteria problem in which the forecast depends not only on the locomotive driver. This subsystem has a limitation associated with the ability to calculate the vector of recommended violations only in the presence of previously recorded violations. This limitation is removed with the help of other approaches used in the considered intellectual system.

VI. REFLECTION

As a result of the study, a model for developing recommendations was built depending on the violations committed by the locomotive driver. With this approach, the problem of "cold start" arises. If the locomotive driver had no violations before, then it is quite difficult to predict a possible violation in the future, since the main feature is what violations were previously committed by the locomotive driver.

TABLE II. THE MOST FREQUENTLY PREDICTED VIOLATIONS OF LOCOMOTIVE DRIVERS

Violation code	Violation name	Group of violations	Number of mentions in the predicted vector
04.14	«There is no overpressure in position 1 before departure»	«Violations of brake control»	23 892
47.34	«violation or absence of checking the auxiliary crane at a speed of 3-5 km per h»	«Other violations in the work of foot-plate staff (crew) »	22 802
43.7	«Violations of the technology for switching on and checking the operation of devices for measuring motion parameters, an integrated locomotive safety device and a complex of information support for an automatic brake control system»	«Violations of the driver during the operation of locomotive safety devices»	22 225
17.3	«Violation or absence of checking the braking equipment when changing the locomotive crew without uncoupling the locomotive from the train or single locomotive»	«Violation of checks of brake equipment during acceptance and delivery of a locomotive »	19 907

The constructed model for the formation of recommendations is the basis for the subsystem of recommendations, which refers to an intelligent system for a comprehensive assessment of the probability of violations by the locomotive driver and the formation of recommendations depending on previously committed violations.

In the future, it is planned to carry out the following works, based on the results presented in this article:

– modernization and revision of the decision rule based on the developed mode;

– implementation of the research results in the information system of JSC «Russian Railways»;

– checking results in the information system JSC «Russian Railways».

The use of an intelligent system for a comprehensive assessment of the probability of a violation by the locomotive driver and making recommendations depending on the previous violations will allow predicting possible violations that the locomotive driver may commit, and improve the reliability of railroads rolling stock safety systems by purposefully managing the human factor. The scientific novelty of the research consists of creating the intellectual system structure, studying of correlation between groups of violations, choice of neural network construction methods and testing of hypothesis about the applicability of the selected

criteria. These tasks were solved for the first time for railway transport.

REFERENCES

[1] Gouriveau R, Medjaher K, Zerhouni N. From prognostics and health systems management to predictive maintenance 1: monitoring and prognostics. 2016.

[2] M.A. Djeziri, S. Benmoussa, R. Sanchez, "Hybrid method for remaining useful life prediction in wind turbine systems", Renew. Energy, pp. 173-187, 2017.

[3] S. Yin, S. Ding, X. Xie and H. Luo, "A review on basic data-driven approaches for industrial process monitoring", IEEE Trans. Ind. Electron., vol. 61, no. 11, pp. 6418-6428, Nov. 2014.

[4] G. Karagiannis, S. Olsen, K. Pedersen, "Deep Learning for Detection of Railway Signs and Signals", In Science and Information Conference; Springer: Berlin/Heidelberg, Germany, pp. 1-15, 2019.

[5] N.Karthick, R.Nagarajan, S.Suresh and R.Prabhu, "Implementation of Railway Track Crack Detection and Protection", International Journal Of Engineering And Computer Science (IJECS), vol. 6, issue 5, pp. 21476-21481, May 2017.

[6] D. Agudo, Á. Sánchez, J.F. Vélez, A. Belén Moreno, "Real-time railway speed limit sign recognition from video sequences", In Proceedings of the International Conference on Systems, Signals, and Image Processing, Bratislava, Slovakia, 23-25 May 2016.

[7] Cernazanu-Glavan and M. Marcu, "Anomaly Detection Using Power Signature of Consumer Electrical Devices", Advanced in Electrical and Computer Engineering, vol. 15, no. 1, pp. 89-94, 2015.

[8] F. Xue, W. Yan, N. Roddy and A. Varma, "Operational data based anomaly detection for locomotive diagnostics", International Conference on Machine Learning; Models Technologies and Applications, pp. 236-241, 2006

[9] I. B. Shubinsky, A. M. Zamyshlyaev, "Main scientific and practical results of URRAN system development", Zheleznodorozhnyi transport, vol. 10, pp. 23-28, 2012.

[10] V. A. Gapanovich, A. M. Zamyshlyaev, I. B. Shubinsky, "Some issues of resource and risk management on railway transport based on the condition of operational dependability and safety of facilities and processes (URRAN project)", Dependability, vol. 1, pp. 2-8, 2011.

[11] I. B. Shubinsky, A. M. Zamyshlyaev, "Main scientic and practical results of URRAN system development", Zheleznodorozhnyi transport, vol. 10, pp. 23-28, 2012.

[12] K. Suzuki, F. Li, S. Sone and K. Doi, "Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network", IEEE Trans. Med. Imag., vol. 24, no. 9, pp. 1138-1150, Sep. 2005.

[13] K. Nakamura, H. Yoshida, R. Engelmann, H. MacMahon, S. Katsuragawa, T. Ishida, et al., "Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks", Radiology, vol. 214, pp. 823-830, 2000

[14] R. E. Turkson, E. Y. Baagyere and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness", 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), pp. 1-7, Sept 2016.

[15] S. V. Mikoni, Theory of managerial decision-making, Saint Petersburg: LAN, 2015.

[16] S. V. Mikoni, B. V. Sokolov, R. M. Yusupov, Qualimetry of models and polymodel complexes, Moscow: RAS, 2018.

[17] B. Kirwan "The validation of three human reliability quantification techniques—THERP, HEART and JHEDI: Part 1—technique descriptions and validation issues", Applied ergonomics. vol. 27, №. 6, pp. 359-373, 1996.

[18] W. Wang, X. Liu, Y. Qin, "A modified HEART method with FANP for human error assessment in high-speed railway dispatching tasks", International Journal of Industrial Ergonomics, vol. 67, pp. 242-258, 2018.

[19] Digital railway: concept of implementation of a complex scientific and technical project, Russia, JSC «RZD», 2017.

[20] V. G. Sidorenko, M. A. Kulagin, "Qualification of drivers as a factor of increasing the reliability of electric rolling stock", Transport science and technology, №4, pp. 70-76, 2018.

[21] V. G. Sidorenko, M. A. Kulagin, "Transport workers activities analysis using an artificial neural network", Proceedings of the Third International Scientific Conference "Intelligent Information Technologies for Industry", "Advances in Intelligent Systems and Computing", vol. 2, pp. 308-316, 2018.

[22] M. A. Kulagin, V. G. Sidorenko, "An approach to forming a driver's performance rating using various comparison metrics", Transport electronics and electrical equipment, № 1, pp. 14-17, 2018.

[23] V. I. Apattsev, V. A. Aksenov, D. L. Raenok, A.M. Zavyalov "Main directions of improving the system of personnel training that ensures the safety of production processes", Science and technology of transport, №1, pp. 93-97, 2014.

[24] V. I. Apattsev, A.M. Zavyalov, I. N. Sinyakina, Yu. V. Zavyalova, E. V. Grishina "Ensuring the safety of train traffic on the basis of reducing the influence of the human factor", Science and technology of transport, №2, pp. 75-78, 2014.

[25] I. B. Shubinsky, A. M. Zamyshlyaev "Topological Semi-Markov Method For Calculation Of Stationary Parameters Of Reliability And Functional Safety Of Technical Systems", Reliability: Theory & Applications, .pp. 12-22, 2012.

[26] G. A. Evstafiev, "Fuzzy cognitive maps in relation to information security risk management", Izvestia of the southern Federal University. Technical science, vol. 100, №11, pp. 45-52, 2009.

[27] P. Jain, P. Netrapalli and S. Sanghavi, "Low-rank matrix completion using alternating minimization", Proc. 45th Annu. ACM Symp. Theory Comput., pp. 665-674, 2013.

[28] Powers DMW. Evaluation: From precision, recall and F-factor to ROC, informedness, markedness correlation. J Mach Learn Technol. pp. 37-63, 2011.

Method for Assessing Probabilistic Reliability Estimation and Safety of Railway Automation Systems Redundant Structures

Dmitry S. Markov
D. Sc., Associate Professor at Department of Automation and Remote Control on Railways, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
mds1945@yandex.ru

Oleg A. Nasedkin
D. Sc., Associate Professor at Department of Automation and Remote Control on Railways, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
Nasedkin@crtc.spb.ru

Alexander D. Manakov
D. Sc., Professor at Department of Automation and Remote Control on Railways, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
manakoff_2@mail.ru

Michael N. Vasilenko
D. Sc., Professor at Department of Automation and Remote Control on Railways, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
vasilenko.m.n@gmail.com

Alexey G. Kotenko,
D. Sc., Chief of Department of Management of Maintenance Works, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
algenko@gmail.com

Vladimir L. Belozerov,
D. Sc., Professor at Department of Transport Economics Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
v.belozerov@mail.ru

Abstract—The article proposes the synthesis features of a modelling algorithm. This algorithm is used to probabilistic reliability estimation and safety of railway automation and remote-control devices. It is used at the stages of development and operational safety proof. To solve the problem, it is proposed to use the simulation method. A direct process simulation model that reproduces the following processes: failure processes, failure detection processes in a system component, system recovery processes. The article describes the following refined definitions: obscure failure, protective failure, dangerous failure, states for the main redundant structures of railway automation microprocessor systems. In the article a formulates the principles and requirements for software tools for building a simulation model. It is proposed to use the GPSS World system as a software tool. Considering the peculiarities of GPSS World and on the formulated principles basis, a modelling algorithm has been designed. The obtained algorithm provides the simplicity of the GPSS-model implementation, which makes it possible to estimate the rates of obscure, protective, and dangerous failures in redundant structures. And similar structures are used to develop the railway automation systems.

Keywords—railway automation and remote-control systems, reliability and safety index, simulation model, obscure and protective failures, dangerous failures, modelling algorithm.

I. Introduction

The task of assessing compliance with the safety requirements of railway infrastructure facilities is solved by certification or declaration. It is assumed that there is evidence for products in the form of "Safety proof of the railway objects" [1 - 6]. One of the components of the safety proof process is the determination of object probabilistic reliability and safety. Besides, it is necessary to verify their compliance with the regulatory value given in the reference documentation or design and technical documentation for the required type of product.

978-1-7281-9899-6/20/$31.00 ©2020

Since the dangerous failures rates values are very small, tests to determine the object probabilistic reliability and safety are practically impossible. The existing methods for assessing the object probabilistic reliability and safety are based on calculation methods using the mathematical apparatus of the reliability theory. However, analytical models require rather strict restrictions on the complexity of the structure of modern microprocessor systems, the probabilistic characteristics of failure processes, detection processes and restorative function processes of railway automation and remote-control devices and subsystems [7]. Overcoming these limitations is achieved using simulation methods. In [8, 9], formalization methods of the railway automation simulation models functioning as complex queuing systems are proposed.

This paper proposes the development of simulation in terms of the processes occurring in the systems of railway automation and remote-control [10 - 12]. The knowledge area related to the processes of failure and restoration of elements and railway automation devices is considered. To solve this problem, an approach based on direct simulation modelling of failure, detection and recovery processes is used. On this basis, a calculated assessment of the railway automation and remote-control systems object probabilistic reliability and safety with a redundant structure or a combination of states of elements that violate the operational or protective state of systems is carried out. The object probabilistic indicators of reliability and safety list include the following “primary” indicators: collective failure rate λ_f, protective failures rate λ_{pf}, dangerous failures rate λ_h, dangerous failures rate $\dot{\omega}_f$, $\dot{\omega}_{pf}$, $\dot{\omega}_h$ of the corresponding failure flow. Applying these indexes,

railway automation and remote-control systems index is calculated as recoverable systems: mean time between failures (*MTBF*), operating time between failures (*OTBF*), probability of no-failure operation $P_{r(t)}$, failure probability $Q_{f(t)}$ for protective and dangerous failures, dependability K_g, availability factor K_{og}, availability factor K_{tu} and safety factor K_s. Individual devices and functional modules are non-recoverable and are characterized by the following probabilistic indicators of reliability and safety: the probability of no-failure operation $P_{r(t)}$; mean time to failure, failure probability $Q_{f(t)}$ also for protective and dangerous failures.

II. The Main Provisions of Railway Automation Systems Reliability and Safety Simulation Modelling

A. Conceptual construct

Let formulate the key concepts in the subject area under consideration.

Definition 1. **Simulation model of the railway automation system [device]:** software analogue like the simulated system [device] concerning the purpose of modelling.

Note: the simulated system can be both real and virtual at the stage of development and safety proof.

Definition 2. **Direct simulation model:** a simulation program where the correspondence of the model virtual objects with real or virtual objects of the modelled system [device] is explicitly established, considering their interconnections.

Definition 3. **Process simulation model:** a simulator that reproduces the processes performed by the system [device] and the processes occurring in it in model time.

Definition 4. **The railway automation system [device] reliability and safety simulation model:** this is a direct process simulator that reproduces discrete processes of failures, element failures detection and the system [device] recovery to probabilistic reliability estimation and safety of the simulated system [device].

Definition 5. **Simulation experiment with a reliability and safety simulation model:** a single run of the simulation model to obtain a point estimate of reliability and safety probabilistic indicators with a given structure, probabilistic values and temporal characteristics of failure processes, failures detection, recovery of the system [device] elements operational state included in the model.

Definition 6. **Simulation experiments planning:** organizing a sequence of simulation experiments that is rational in terms of the computer time cost to determine the dependence of the probabilistic indicators of reliability and safety on the probabilistic and temporal characteristics of failure processes, detect failures, restore the operational state of the system [device] elements and ensure the required quality of simulation results.

Definition 7. **A simulation experiments series with a simulation model of reliability and safety:** a sequence of simulation experiments carried out following a given plan to obtain a set of point estimates of probabilistic reliability estimation and safety, sufficient to determine the dependence of probabilistic reliability estimation and safety on the probabilistic-temporal characteristics of failure processes, failure detection, recovery of all system [device] elements working state included in the model.

Definition 8. **A simulation session with a simulation model of reliability and safety:** a set of simulation experiments series performed under strategic and tactical plans for the whole set of probabilistic reliability estimation and safety, a set of dependencies of probabilistic reliability estimation and safety on various factors to ensure the completeness of the assessment of the simulated system [device] by probabilistic reliability estimation and safety.

The use of a reliability and safety direct process simulation model to assess the probabilistic indicators of reliability and safety requires clarification of the conceptual apparatus for the events and states of redundant structures of railway automation microprocessor systems, taking into account the means of detecting failures, primarily according to their time characteristics. By this, we will introduce several definitions, considering the redundant structures of railway automation microprocessor-based systems as objects of failure processes simulation modelling, failure detection and restoration of an operational state. It should be noted that the same elements or devices failure can bring the system into different states depending on the time of occurrence. Definitions are introduced for the main structures that have found the practical application of railway automation microprocessor-based systems, namely
2^2 (2oo2), 2^3 (2oo3), 2^2v2^2 (1oo2D). Diagnostic tools and interface devices with control objects are considered reliable. In the reliability and safety simulation model, they are considered as separate devices or subsystems and, if necessary, are included in the model with their probabilistic-temporal characteristics.

Definition 9. **The railway automation system [device] operative state with redundant structure:** the railway automation system [device] state in which the state of all elements, functional blocks and sub-systems are operational per the requirements of technical documentation.

Note: in the context of this work, scratches on the case do not render the railway automation systems out of order.

Definition 10. **The railway automation system [device] operable state with a redundant structure:** this state includes an operative state and a faulty state in which the railway automation system [device] with a redundant structure performs the required functions under the technical documentation when individual elements, functional blocks, subsystems are inoperative.

Note: the railway automation system [device] operable state with a redundant structure in a faulty state is ensured by its reserve capabilities.

In [7, 8], the conceptual apparatus in the railway automation reliability and safety field is formulated, which for simulation experiments must be specified considering the redundant structures of technical means used in practice.

Definition 11. **Protective failure of structure 2oo2:** the forced transfer of the structure to an inoperative state after a failure is detected in one of the information processing channels.

Definition 12. **The protective state of structure 2oo2:** inoperative state from the moment a protective failure occurs until the system is restored to an operable state.

Definition 13. **Dangerous failure of the 2oo2 structure:** an event associated with the occurrence of a failure structure in the computational channels in the time interval from its occurrence to the detection of the previous failure.

Definition 14. **Obscure failure of structure 2oo3:** one of the channels failures in the structure.

Definition 15. **The fault state of structure 2oo3:** the structure operable state from the moment of occurrence of an obscure failure until its elimination.

Definition 16. **Protective failure of the 2oo3 structure:** the forced transfer of the structure to an inoperative state when a failure is detected in one of the information processing channels after its reconfiguration into the 2oo2 structure.

Definition 17. **The protective state of structure 2oo3:** the disabled state of the structure from the moment the protective failure occurs to the restoration of the upstate of at least one of the failed channels.

Definition 18. **Dangerous failure of structure 2oo3:** an event associated with the occurrence of a failure structure in the computational channels in the time interval from its occurrence to the detection of the previous failure.

Definition 19. **Obscure failure of structure 1oo2D:** failure of one channel in an active or standby set.

Definition 20. **The fault state of structure 1oo2D:** the operational state of the system from the moment a masked fault occurs until it is rectified.

Definition 21. **Protective failure of structure 1oo2D:** the forced transfer of the structure to a disabled state when a failure occurs in a running set during the time from detection to the elimination of a failure in a restored set.

Definition 22. **The protective state of structure 1oo2D:** the disabled state of the system from the moment of the occurrence of the protective failure to the restoration of the upstate of one of the sets of the 1oo2D structure.

Definition 23. **Obscure of structure 1oo2D:** a failure occurring in a running or backup system set from the time of occurrence to the detection of a previous failure in the same set.

B. Design Principles for a Reliability and Safety Simulation Model

1. The simulation model for assessing the reliability and safety indicators of railway automation systems is developed as a direct process simulation model of reliability and safety by the above definitions.

2. The railway automation reliability and safety simulation model is the object of a simulation experiments sets, by this, all probabilistic indicators of reliability and safety, and first of all the "primary" λ_f, λ_{pf}, λ_h, $\dot{\omega}_f$, $\dot{\omega}_{pf}$, $\dot{\omega}_h$ are evaluated not analytically, but by mathematical statistics methods.

3. Reducing the computer time cost for performing simulation experiments sets for the statistical assessment of probabilistic reliability estimation and safety, taking into account the low intensity of railway automation systems dangerous failures flows.

4. According to the technique of the conducting experiment, it is advisable to use the concept of the failures flow parameter, which makes it possible to carry out a simulation experiment in the form of one long realization of the failures processes, detecting failures and recover railway automation (modelling one system for a long time). This approach significantly simplifies the procedures of the simulation experiment with the set of implementations when simulating recoverable and non-recoverable systems in terms of operating time to failure (modelling a set of devices over a relatively short time interval). For recoverable systems, the time between failures can be considered as mean time to failure. Simplification of the modelling procedures for one long implementation is achieved by the simplicity of solving the problem of initial conditions, stopping the modelling process, and the absence of the need to organize a restart of the model to implement short runs.

5. The detecting failures main tools of railway automatics microprocessor-based systems are cyclic diagnostics with a period τ_d, which is negligible compared to the mean time between failures. It is impossible to clock the simulation model of reliability and safety with the period τ_d due to the prohibitive time of the simulation experiment implementation. In contrast to a real system, the model "knows" the moments of failure, which allows the model to be clocked by the operating time between failures. Then, at the failure occurrence moment, time delays $\tau^o = \tau_d + \tau_{det}$ and τ^r are sequentially started, where τ^o is the time of failure detection, τ_{det} is the time of performing operations to detect a failure by τ_d, and τ^r is the recovery time and the procedures for switching on the device. Besides, it takes into account the operation of functional diagnostics, safe comparison circuits, etc.

6. Information about the simulated railway automation system for the elements set included in the simulation model of reliability and safety and their probabilistic-time characteristics are entered into the model in the form of initial data.

7. The simulated railway automation system structure from the viewpoint of reliability and safety is represented in the functions logical form of the states of elements that describe the state of the system, for which the probabilistic indicators of reliability and safety are assessed.

C. Requirements for a simulation model of reliability and safety

The introduced conceptual apparatus (clause *A*) and development principles (1-7) (clause *B*) make it possible to formulate requirements for a simulation model of reliability and safety:

1. The reliability and safety simulation model should be implemented as a complex, including the following programs: the actual simulation model of railway automation reliability and safety; an interactive subsystem that implements the user interface with the model, including a module for setting up a reliability and safety simulation model for the structure of a railway automation simulated system [device], a module for setting up a simulation model for a parametric description of failure processes, detecting element failures and restoring an operable (serviceable) state of a modelled railway automation systems [devices], a module for setting plans for a series of imitation experiments, a module for setting up

the presentation and output of the results of a series of imitation experiments.

2. Railway automation elements (devices, modules, subsystems) should be explicitly represented in the reliability and safety simulation model as dynamic objects that generate the failures processes, their detection and restoration of an operational state.

3. An element of the modelled system in the reliability and safety simulation model can be in the following states:

- the component is operational;
- a component in a state of failure of a certain type from a set of user-specified failure types, respectively, different failure states of an element, as many as specified failure types;
- state of failure of the component until a failure is detected;
- state of failure of the component until the restoration of the upstate of the system [device].

4. Failures processes, elements failures detection and operational state recover of railway automation systems [devices] should be carried out in the simulation model of reliability and safety at the probabilistic-temporal level without disclosing detection and recovery technologies in the model.

5. The core of the reliability and safety simulation model should be a module for determining the coincidence of the states of elements specified by the user, i.e. those states of the modelled system [device] whose intensity (probability) is estimated in a series of simulation experiments. Matching should be done under the structure of the railway automation system (direct model).

6. The reliability and safety simulation model should not impose restrictions on the structure and probabilistic-temporal characteristics of failure processes, failure detection and recovery of devices and systems of railway automation inherent in analytical reliability models.

7. Ensuring the independence of failure processes, failure detection and recovery for various elements should be ensured by their generation using various generators of uniformly distributed random numbers on the interval [0,1].

8. The adequacy of the reliability and safety simulation model should be confirmed by verification and validation procedures for all stages of the life cycle, expert assessments, calculations of analytical models using a simplified formalized scheme and statistical data on the actual operation of railway automation (as test examples).

9. The probabilistic-temporal characteristics of the failure of the elements flows should be set in the module for setting up the reliability and safety simulation model for the parametric description of the modelled system using the fastest tabular form of description of random variables.

10. The times for detecting failures of elements and recover the operational state of the system [device] are set by the user and can vary from deterministic values to random values generated similarly to the probabilistic-time characteristics of failure flows.

11. The reliability and safety simulation model should include means of organizing tactical and strategic plans for a series of simulation experiments, generating reports of various forms per the list of estimated probabilistic indicators of reliability and safety.

12. Completion of simulation experiments should be carried out to meet the requirements of a given value of the confidence interval at a given confidence probability or the convergence of the probabilistic characteristics of the rarest situations (combinations of states of elements included in the modelled system) from the whole set of situations, the probabilistic indicators of reliability and safety of which are evaluated in this experiment.

D. Choosing a software tool for implementing a reliability and safety simulation model

Under clause C, software tools must meet the following requirements:

- keeping the model time by events;

- practically unlimited number of generators of uniformly distributed random variables on the interval [0-1];

- the random variables generators presence with different distribution laws, the most common in research on the reliability of technical systems;

- reproduction of discrete processes;

- conducting parallel discrete processes;

- availability of means for representing elements of systems (devices);

- availability of means for determining the state of model objects and their combinations;

- availability of computational capabilities used directly in the simulation process;

- availability of built-in tools for collecting and storing simulation results;

- availability of built-in tools for preliminary processing of statistical data;

- the possibility of organizing by control teams of an automatic mode of performing a series of simulation experiments;

- the presence of built-in tools for syntactic and logical control of the program text and the modelling process;

- the ability to visually control the progress of the modelling process;

- the ability to interact with other software tools to organize the most comfortable dialogue subsystem in a specific subject area.

Almost all the requirements are met by the GPSS World software tools [13], the authors have considerable experience with it. The main disadvantage of GPSS World, from developing a reliability and safety simulation model, is its rather weak dialogue capabilities. The comfortable interface "user-reliability and safety simulation model" development is quite simply implemented employing interaction between GPSS World and other software tools.

III. MODELLING ALGORITHM

The modelling algorithm was developed considering the subsequent implementation of the reliability and safety simulation model as a GPSS-model and under the provisions of clause 2. The main difference of the proposed algorithm from the known reliability simulation models of technical systems in the GPSS environment [14] is the use of a simulated system of dynamic objects - transactions - to represent devices. This allows, in contrast to the use of hardware GPSS objects

(single-channel devices, memories, logical keys), to synthesize a universal algorithm and a GPSS-program, in which the composition and structure of the system under study are set at the level of the initial data without changing the text of the simulating program. The modelling algorithm of reliability and safety simulation model is presented in the form of the algorithm logic diagram [15, 16]:

$$MA=O_0\ O_1\ O_2\ \downarrow^3 O_3\ O_4\ q_1\uparrow^1\ O_5\ q_2\uparrow^1\ O_6\ O_7\ K$$
$$\downarrow^1 O_8\ O_9\ O_{10}\ q_3\uparrow^2\ O_{11}\ \downarrow^2 O_{12}\ O_{13}\ \dot{\omega}\uparrow^3$$

Here is a description of the modelling algorithm:

Operator O_0 – the start of algorithm execution.

Operator O_1 – writing initial data to GPSS World memory objects. The initial data for the modelling algorithm are:

- J – number of devices included in the reliability and safety simulation model;

- λ_j - failure rate, $OTBF_j$ operating time between failures of $j^{\text{-th}}$ component $j=\overline{1,J}$;

- FN – unit table function of the distribution of a random variable;

- τ_j^{o} failure detection time, τ_j^{r} restoration time of $j^{\text{-th}}$ device;

- K, M - the number of combinations of device states from the set J, corresponding to dangerous conditions and protective conditions of the system, respectively;

- $BV^{\text{o}}_{k,j}$, $BV^{3}_{m,j}$, $k=\overline{1,K}$, $m=\overline{1,M}$ - Boolean functions describing K dangerous conditions and M protective conditions of the system;

- X^{z}_{h} – memory cell index into which the value of the number of dangerous failures of the system is written, upon reaching which the simulation process is stopped.

Operator O_2 performs transaction Tr_j generation and assigns values $OTBF_j$, τ_j^{o}, τ_j^{r}, k_j, m_j to each $j^{\text{-th}}$ transaction. Thus, a lot of transactions Tr_j; $j=\overline{1,J}$ are created, representing devices of the system under study with their properties in the model. Obviously, in this case, when researching a specific system, it is not the modelling algorithm and GPSS program that are changed, but the number of processed transactions that correspond to the number of devices included in the reliability and safety simulation model.

Operator O_3 delays each transaction Tr_j for a random amount of time between failures of the $j^{\text{-th}}$ device, determined by the value of $OTBF_j$ and the unit distribution function FN. It should be noted that this approach is used for one-parameter functions, for example, in the exponential distribution or Rayleigh distribution, most common in studies of the reliability of technical systems. For two-parameter distribution functions (for example, normal distribution) of time between failures, the corresponding unit functions FN is also used, and the values of their parameters are specified in the initial data (Operator O1).

A timeout of transaction Tr_j means that event occurs – the failure of the $j^{\text{-th}}$ device. Two memory cells are allocated for each device. Cell X_{2*j-1} contains device state from failure till completion τ_j^{o}, and cell X_{2*j} – from completion τ_j^{o} till completion τ_j^{r}.

Operator O_4 fixes failure $X_{2*j-1}:=1$. So in memory cells X_{2*j-1}, $j=\overline{1,J}$ the current state of the system by the presence or absence of failures of its devices is stored. System dangerous conditions, e.g. as defined in 13, 18, 23, according to the current state of devices are determined by the set K of Boolean functions - $BV^{\text{o}}_{k,j}(X_{2*j-1})$, whose arguments are memory cell X_{2*j-1} indices corresponding to devices whose joint failure puts the system in dangerous conditions.

Logical condition q_1 defines the values of those functions $BV^{\text{o}}_{k,j}(X_{2*j-1})$, which may change when the value of the memory cell X_{2*j-1} changes by transaction Tr_j, $q_1=1$, if $BV^{\text{o}}_{k,j}(X_{2*j-1})=1$ and $q_1=0$ otherwise.

Operator O_5 when $q_1=1$ captures the current number of dangerous conditions $X_h:=X_h+1$.

Logical condition q_2 compares the current number of dangerous conditions X_h with a given X^z_h for stopping the simulation experiment process. If $X_h > X^z_h$ $q_2=1$, otherwise $q_2=0$.

Operator O_6 when $q_2=1$ calculates the rate of the dangerous condition $\lambda_h = X_h/T_{mod}$, where T_{mod} the value of the absolute model time at the end of the simulation experiment.

Operator O_7 calculates protective failure rate $\lambda_p = X_p/T_{mod}$.

Operator K performs the withdrawal of active transactions from simulated reliability and safety model and the completion of this simulation experiment.

Operator O_8 when $q_1=0$ or $q_2=0$ delays transaction Tr_j for failure detection time τ_j^{o}.

Operator O_9 at the end of τ_j^{o} admits the failure as detected and $X_{2*j-1}:=0$, thus, in this implementation, this device is excluded from the analysis of the system dangerous conditions.

Operator O_{10} $X_{2*j}:=1$ fixes the recover beginning of the device. Protective states of the system, for example, as defined in 12, 17, 22, according to the current state of devices are determined by a set M of Boolean functions – $BV^{3}_{m,}(X_{2*j})$, whose arguments are indices of memory cells X_{2*j} corresponding to devices, the joint failure of which puts the system in a protective state.

Logical condition q_3 defines the values of those functions $BV^{3}_{m,j}(X_{2*j})$, which may change when the value of the memory cell X_{2*j} changes by transaction Tr_j, $q_3=1$ if $BV^{3}_{m,j}(X_{2*j})=1$, $q_3=0$ otherwise.

Operator O_{11} is performed when $q_3=1$ $X_3:=X_3+1$ captures the current number of protective states.

Operator O_{12} delays transaction Tr_j for the restoration time of $j^{\text{-th}}$ device τ_j^{r}.

Operator O_{13} at the end of τ_j^{r} admits the device as restored $X_{2*j}:=0$, thus, in this implementation, this device is excluded from the analysis for the protective state of the system.

Further, the transaction Tr_j by the identically false logical condition $\dot{\omega}\uparrow^3$ transmitted to the operator O_3, thereby organizing a cycle. Thus, all J transactions representing J devices of the system included in the simulated reliability and safety model are "rotated" in the model until the end of the simulation experiment according to condition q_2.

The presented algorithm has been tested on several examples and has shown its performance.

IV. Conclusion

The paper shows the relevance of the reliability and safety simulation model synthesis with the aim of to assess the probabilistic reliability estimation and safety for proving the railway automation and remote-control modern systems safety at the stages of development and certification for safety. The following results were obtained:

- a set of terms and definitions was formulated in the fault processes simulation modelling field, fault detection and restoration of the operable condition of railway automation devices and systems;

- the reliability and safety simulation model is defined as a direct process model, which explicitly reproduces the processes of failures, their detection and recovery, which removes the limitations inherent in analytical models;

- the terms and definitions of dangerous and protective failures and states for the main redundant structures of railway automatics microprocessor-based systems were clarified, taking into account the time of detection and recovery of operability;

- the reliability and safety simulation model synthesis principles are formulated, the main of which is the representation of system devices by dynamic objects of the model and the provision, on this basis, of the possibility of setting the properties of the simulated system at the initial data level without changing the text of the simulating program, i.e., the universality of the model for a given subject area;

- a set of requirements for the tool has been determined, which made it possible to choose a system for modelling discrete processes in continuous time GPSS World for the implementation of a simulation model of reliability and safety;

- a modelling algorithm has been developed based on the formulated principles and taking into account the subsequent implementation of a reliability and safety simulation model in the GPSS World environment, which ensures the universality of the model and a description of the system structure concerning the assessment of probabilistic reliability estimation and safety by a set of Boolean functions.

Perspectives:

- synthesis of reliability and safety simulation model based on the proposed modelling algorithm in the GPSS World environment with the development of the means of the dialogue subsystem following subparagraph 1 of paragraph C;

- tactical and strategic plans development for conducting a series of simulation experiments to assess the probabilistic reliability estimation and safety of redundant structures;

- software selection for calculating regression models of dependences of “primary” and complex probabilistic reliability estimation and safety on the parameters of the studied system of railway automation;

- development of verification and validation procedures; simulation model of reliability and safety;

- execution of control sets experiments under the developed plans;

- verification and validation of the simulation model of reliability and safety using the results of control experiments;

- planning and assessment of probabilistic indicators of reliability and safety by carrying out a series of simulation experiments for real microprocessor-based systems of railway automation.

References

[1] Gavzov D.V., Sapozhnikov V.V., Sapozhnikov Vl.V. Methods for providing safety in discrete systems // Autom. Remote Control, 55:8 (1994), 1085–1122.

[2] Smith D.J., Simpson K.G.L. Functional safety: A Straightforward Guide to IEC 61508 and Related Standards. – Butterworth-Heinemann; 1st edition (June 26, 2001), 208 p.

[3] Bukowski J.V., van Beurden I. Impact of proof test effectiveness on safety instrumented system performance // 2009 Annual Reliability and Maintainability Symposium, Fort Worth, TX, USA, 26-29 Jan. 2009, DOI: 10.1109/RAMS.2009.4914668.

[4] Smith D. J., Simpson K.G.L. Safety Critical Systems Handbook. A Straightforward Guide to Functional Safety, IEC 61508 and Related Standards, Including Process IEC 61511 and Machinery IEC 62061 and ISO 13849 // Oxford, UK, Elsevier Ltd, 2010, 270 p.

[5] Julserccwong A., Thepmanee T. Design and Implementation of Functional Safety for Repairable Systems // 2018 57th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Nara, Japan, 11-14 Sept. 2018, DOI: 10.23919/SICE.2018.8492691.

[6] Belishkina T.A. Features of confirmation of compliance with railway safety requirements during the transition period after the adoption of technical regulations of the customs union [Osobennosti podtverzhdeniia sootvetstviia trebovaniiam bezopasnosti zheleznodorozhnoi` v perehodny`i` period posle priniatiia tekhnicheskikh reglamentov tamozhennogo soiuza] // Automation on transport. – 2016. – Vol. 2. – №2. – P. 23-27. (in Russian)

[7] Shubinskii I.B. Reliable fault tolerant information systems. Synthesis methods [Nadezhny`e otkazoustoi`chivy`e informatcionny`e sistemy`. Metody` sinteza]. – Moscow.: «Reliability magazine», 2016, 546 p. (in Russian)

[8] Markov D.S., Bulavskii` P.E. The matrix method of formalizing simulation models of complex mass service systems [Matrichny`i` metod formalizatcii imitatcionny`kh modelei` slozhny`kh sistem massovogo obsluzhivaniia] // PSTU proceedings, 2010, №4, P. 186-195. (in Russian)

[9] Markov D.S., Lykov A.A. The method of formalization of simulation models of technological processes in the economy of automation and remote control in railway transport [Metod formalizatcii imitatcionny`kh modelei` tekhnologicheskikh protcessov v hoziai`stve avtomatiki i telemehaniki na zheleznodorozhnom transporte] // PSTU proceedings, 2012, №1, P. 23-38. (in Russian)

[10] Antoni M. Formal validation method for computerized railway interlocking systems // 2009 International Conference on Computers & Industrial Engineering, Troyes, France, 6-9 July 2009, DOI: 10.1109/ICCIE.2009.5223968.

[11] Efanov D., Lykov A., Osadchy G. Testing of Relay-Contact Circuits of Railway Signalling and Interlocking // Proceedings of 15th IEEE East-West Design & Test Symposium (EWDTS`2017), Novi Sad, Serbia, September 29 – October 2, 2017, pp. 242-248, doi: 10.1109/EWDTS.2017.8110095.

[12] Theeg G., Vlasenko S. Railway Signalling & Interlocking: 2nd Edition. – Germany, Hamburg: PMC Media House GmbH, 2018, 458 p.

[13] Computer Simulation: http://www.minutemansoftware.com/simulation.htm

[14] Zhernovyi Iu.V. Simulation models for calculating the reliability of systems. Using GPSS World. [Imitatcionny`e modeli rascheta nadezhnosti sistem. Ispol`zovanie GPSS World]. – Kiyv: LAP LAMBERT Academic Publishing, 2018, 96 p. (in Russian).

[15] Lazarev V.G., Piel E.I. Synthesis of control machines [Sintez upravliaiushchikh avtomatov]. – Moscow.: Energiia, 1978, 408 p. (in Russian).

[16] Baratov D.K., Aripov N.M., Ruziev D.Kh. Formalized Methods of Analysis and Synthesis of Electronic Document Management of Technical Documentation // 2019 IEEE East-West Design & Test Symposium (EWDTS), Batumi, Georgia, 13-16 Sept. 2019, DOI: 10.1109/EWDTS.2019.8884415.

Bio-inspired Approach to Microwave Circuit Design

Vladislav Ivanovich Danilchenko
Computer-aided design of department, postgraduate

Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education "Southern Federal University".

Taganrog, Russia.
vdanilchenko@sfedu.ru

Yevgenia Vladimirovna Danilchenko
Computer-aided design of department, postgraduate

Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education "Southern Federal University".

Taganrog, Russia.
lipkina@sfedu.ru

Viktor Mikhailovich Kureichik
Computer-aided design of department, postgraduate

Federal State-Owned Autonomy Educational Establishment of Higher Vocational Education "Southern Federal University".

Taganrog, Russia.
vmkureychik@sfedu.ru

***Abstract*— The article describes a bio-inspired approach to the structural and parametric design of microwave circuits based on a genetic algorithm (GA), which permits algorithmic environment in the field of genetic search for solving NP complete problems, in particular, structural-parametric synthesis of a microwave amplifier. The article is aimed at finding ways of structure-parametric synthesis of microwave modules based on a bio-inspired theory. Scientific novelty of the research lies in the development of a modified genetic algorithm for a bio-inspired automated structural and parametric synthesis of microwave modules. The problem statement in the paper is as follows: to optimize the synthesis of passive and active microwave circuits by using a modified GA. The fundamental distinction of a new approach from the known is the use of new modified genetic structures in bio-inspired automatic structural and parametric synthesis, moreover, a new method for calculating a microwave amplifier based on modified GA is presented in the work. Thus, the problem of creating methods, algorithms and software for automated structural synthesis of microwave modules is currently of special actuality. The solution of the problem will improve the quality characteristics of the designed devices; reduce the time and cost of design and lower the requirements for implementation qualifications.**

***Keywords*— genetic algorithms, graphics and hypergraphs, automated calculations, microwave modules, CAD, circuit diagram, topology.**

I. INTRODUCTION

The problems in the performance of the task of structural-parametric synthesis of passive and active microwave circuits based on classical GAs include a large variety of these circuits that complicates their systematization and development of a universal encoding-decoding algorithm, as well as leads to redundant or unrealizable solutions. According to the paper [1], the problem of structural synthesis of REE refers to nondeterministic polynomial time complete problems [2]. This means that, generally, its solution in principle cannot be found in end-time by any algorithm. To eliminate the problem, revised versions of the GA using prior knowledge about the designed class of devices were developed. He following authors made a significant contribution into the development of that particular direction: L.S. Berstein, G.G. Kazennov, V.P. Koryachko. V.M. Kureichik, I.P. Norenkov, L.A. Rastrigin, G.G. Ryabov, P.I. Sosnin, A.L. Stempkovskiy, L. Goldberg, D. Holland and others.

When designing, electro migration and all parasitic effects affecting the correct functioning of the device should be taken into account (the dependence of the electrical characteristics of the compounds on the purpose of the layers, the resistance of interlayer transfers, cross-coupled interference and noise).

An increase in the number of layers complicates the lithography process, and, consequently, the cost of product development raises. Reduction in the number of layers means an increase in density of the compounds, which, in turn, leads to an increase in power dissipation and deterioration of technical specifications. Constant reduction of geometric dimensions of discrete elements and reduction in the width of the conductors lead to a decrease in current density and distortion of the transmitted signals. Circuits' interval reduction promotes the effect of subtle physical effects such as parasitic capacitance, inductance, and electro migration. Reducing interlayer spacing is limited by the capacities of production technology, but the smaller it is; the more significant is the interlayer capacity. The number of criteria and limitations that must be considered when solving the problem of microwave circuits design is growing. New approaches and algorithms to solve the problem are required. A hybrid algorithm allowing verification directly at the development stage is needed for that kind of problem [3].

II. PROBLEM STATEMENT

Normally, a complex microwave circuit is represented in the form of a certain combination of elements whose equivalent parameters are known. Propagation of a single-mode TEM-wave in transmission lines between individual elements is proposed, which makes possible the use of the following expressions to calculate the complex amplitudes of voltage and current and cross-section x (Fig. 1):

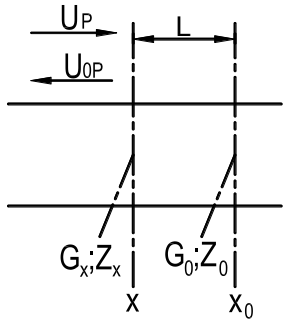


Fig. 1. Cross-section X or calculating the complex amplitudes of voltage and current

$$U(x) = U_{p0}\ \exp[-y_0(x - x_0)] + U_{op0}\ - \exp[y_0(x - x_0)], \quad (1)$$

$$I(x) = \left(U_{p0}/\rho\right)\exp[-y_0(x - x_0)] \left(U_{op0}/\rho\right)\exp[y_0(x - x_0)], \quad (2)$$

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

where U_{p0} и U_{op0} - complex amplitudes of the voltage of the incident and reflected waves at $x = x_0$; ρ – line impedance; $y_0 = \beta_0 + ja_0$ – propagation constant; β_0 – attenuation constant; $a_0 = 2\pi/\lambda_d$ – phase shift constant; λ_d – line wave-length.

When switching on the end of the line of the complex load Z_n according to (1) and (2), the input resistance in the cross-section x

$$Z_{vx} = \frac{U(x)}{I(x)} = \rho \frac{Z_n ch y_0 l + \rho sh y_0 l}{\rho ch y_0 l + Z_n sh y_0 l}, \quad (3)$$

where l – line length,

$$ch(y_0 l) = ch(\beta_0 l)cos(a_0 l) + jsh(\beta_0 l)sin(a_0 l), \quad (4)$$

$$sh(y_0 l) = sh(\beta_0 l)cos(a_0 l) + jch(\beta_0 l)sin(a_0 l). \quad (5)$$

Neglecting the resistive losses in the line ($\beta_0 = 0$) we write expression (3) in the following form

$$Z_{vx} = \rho(Z_n + j\rho tg a_0 l)/(\rho + jZ_n tg a_0 l). \quad (6)$$

The properties of a linear two-terminal network are described with the use of the reflection coefficient (Fig. 2, a),

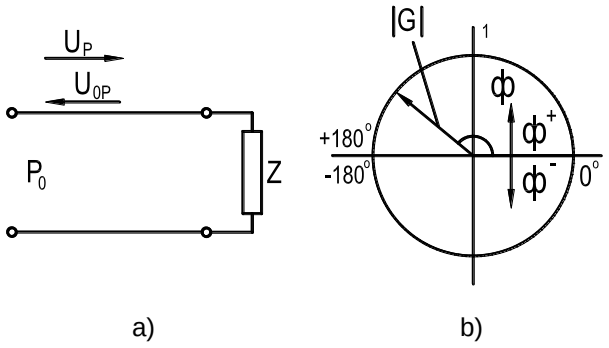


Fig. 2. Reflection coefficient of a two-terminal network

$$G = \frac{U_{op}}{U_p} = \frac{Z - \rho_0}{Z + \rho_0} = |G| exp(j\varphi_g), \quad (7)$$

where ρ_0 – standard value of wave impedance.

To calculate the modulus and phase of the reflection coefficient according to (7) (Fig. 2, b), algorithm 1 is used, in which VBB is the reflection coefficient G, RMT is the module $|G|$, FS is the phase φ_g, in degrees.

Both individual elements of a linear type and the entire microwave device made up of such elements are characterized by a scattering wave matrix (S-parameters) or transmission (T-parameters), connecting the voltage amplitudes of the incident and reflected waves in external transmission lines. To eliminate the normalization operation, it is assumed that all external lines connected to microwave devices have a standard value of wave impedance ρ_0 = 50 Ohm [3, 4]. For a four-terminal network (Fig. 3) in the S-parameter system we have

$$U_{10} = S_{11}U_{1p} + S_{12}U_{1o}, U_{2p} = S_{21}U_{1p} + S_{22}U_{2o}. \quad (8)$$

In the T-parameter system:

$$U_{1p} = T_{11}U_{2p} + T_{12}U_{2o}, U_{1o} = T_{21}U_{1p} + S_{22}U_{2o}. \quad (9)$$

Algorithm 1 – Calculation of the phase of the reflection coefficient

```
1 SUBROUTINE VKF (VBB, RMT, RS)
2 COMPLEX VBB
3 G=180./3.14159265
4 RMT=CABS (VBB)
5 PB=REAL (VBB)
6 QB=AIMAG (VBB)
7 IF (PB) 2, 3, 4
8 FS=G*ATAN (QB/PB)
9 GO TO 8
10 IF (QB) 5, 6
11 FS = -180.+G*ATAN (QB/PB)
12 GO TO 8
13 IF (QB) 7, 8, 9
14 FS = 90.
15 GO TO 8
16 FS = -90.
17 GO TO 8
18 FS = 0.0
19 CONTINUE
20 RETURN
21 END
```

With known T - parameters, S - parameters can be calculated according to algorithm 2.

Algorithm 2 – Calculation of S-parameters

```
1 SUBROUTINE PTS (T11, T12, T21, T22, S11, S12, S21, S22)
2 COMPLEX T11, T12, T21, T22, S11, S12, S21, S22
3 S11=T21/T11
4 S12=T22-T12*T21/T11
5 S21=1./T11
6 S22=-T12/T11
7 RETURN
8 END
```

We distinguish between two transfer coefficients of a four-terminal network by voltage: in relation to the input voltage:

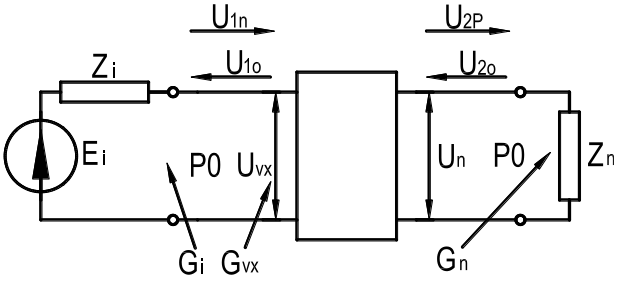


Fig. 3. Example of a four-terminal network for calculation

$K_g(jw) = \frac{U_n}{U_{vx}}$ and at EMF of signal source $K_i(jw) = U_n/E_i$ (Fig. 3).

For the first of them, taking into account (8) and the expressions for the total voltages at the load at the input of the four-terminal network $U_{vx} = U_{1p}(1 + G_{vx})$ we get:

$$K_g(jw) = \frac{U_n}{U_{vx}} = \frac{S_{21}(1+G_n)}{(1-S_{22}G_n)(1+G_{vx})}, \quad (10)$$

где $G_{vx} = S_{11} + S_{12}S_{21}S_g/(1 - S_{22}G_n)$ – reflection coefficient at the input of a four-terminal network. From (10) the phase-frequency characteristic of the circuit:

$$\varphi_g(jw) = arctg\left[\frac{Im\left(K_g(jw)\right)}{Re\left(K_g(jw)\right)}\right]. \quad (11)$$

Taking (8), (9) and the ratio $U_1 = Z_{vx}E_i/(Z_i + Z_{vx})$; $Z_i = \rho_0(1 + G_i)/(1 - G_i)$; $Z_{vx} = \rho_0(1 + G_{vx})/(1 - G_{vx})$ into account for repeated transfer constant we have [5]:

$$K_i(jw) = \frac{U_n}{E_i} = \frac{S_{21}(1+G_n)(1-G_i)}{2(1-S_{22}G_n)(1+G_{vx})}, \quad (12)$$

Power factor of a reactive four-terminal network (Fig. 3):

$$K_p = \frac{\mathrm{P}_n}{P_{ji(nom)}} = \frac{(1-|G_{vx}|^2)(1-|G_i|^2)}{|1-G_{vx}G_i|^2} \leq 1. \quad (13)$$

where P_n – power in load, $P_{ji(nom)} = E_i^2/8Re(Z_i)$ – nominal power of signal source.

The expressions (10-13) allow calculating of amplitude-phase frequency characteristics of four-terminal network and its attenuation $b_3 = 10lgK_{\mathrm{P}}$.

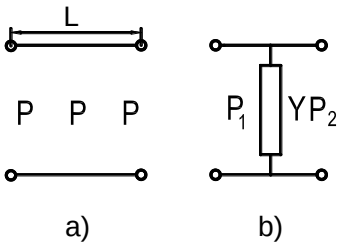


Fig. 4. Compounds of two lines with different wave impedance

For wo most typical microwave four-terminal networks: the line length (Fig. 4, a) and the connection of two lines with different wave impedances (Fig. 4, b), T-parameters are calculated according to algorithm 3, in which the following notation is used: FS - phase line angle $\theta = 2\pi l/\lambda_д$,in radians.

The main features of the proposed programs for the analysis and synthesis of various microwave devices are as follows: in all cases, the designed circuit is calculated directly considering its distributed structure without any equivalent transformations, which increases the accuracy of the final result:

Algorithm 3 – Calculation of T-parameters

```
1 SUBROUTINE TL (FS, T11, T12, T21, T22)
2 COMPLEX T11, T12, T21, T22
3 B=COS (FS)
4 C=SIN (FS)
5 T11=CMPLX (B, C)
6 T12=CMPLX (0., 0.)
7 T21=CMPLX (0., 0.)
8 T22=CMPLX (B, -C)
9 RETURN
10 END
```

- The required circuit characteristic forms the basis of the objective function, low approximation is not applicable;
- There may be limitations of the structural and technological characteristics applied individual elements of the circuit;
- It is possible to take all the heterogeneities in the microwave circuit into account, including those at the junction of individual links;
- It is allowed to connect various additional elements to the device for its articulation with other functional units;
- Frequency-dependent complex resistance may be the load;
- The properties of the synthesized circuit can be evaluated according to several criteria.

All these features are additional advantages of designing microwave devices with the use of genetic algorithms.

III. Hybrid Genetic Algorithm

This paper presents the results of experiments to determine the quality of the solution, the algorithm for generating the starting population, and genetic operators. Based on the research results, the optimal control parameters for a complex of algorithms for the synthesis of passive and active microwave circuits were determined.

For the task of synthesizing passive and active microwave circuits using a modified HA, it is necessary to determine the following criteria: total length of connections; number of vias; total interlayer capacity; the total resistance of the conductors, which determines the time delays. It is necessary to bring all the criteria to a single form or to normalize them. Normalization will ensure that all criteria are equal; they will take values in the range from 0 to 1 [7].

For normalization, the maximum possible length of the connections must be determined. To do this, after creating the initial population, the maximum value of the length of all connections is determined, and after each iteration, it is checked whether a large value was received.

All other criteria, namely the number of vias, the interlayer capacitance and the total resistance of the conductors, are normalized by analogy with the previous criterion.

The currently existing methods of multicriteria optimization can be conditionally divided into two groups [8]. The methods of the first group reduce the multicriteria problem to single-criterion optimization. There are different types of packages. The most common method of convolution of a vector criterion is linear convolution of the form:

$$\Phi(x) = \sum_{i=1}^{m} \alpha_i F_i(x), \alpha_i \geq 0. \quad (14)$$

This type of convolution is used in the genetic algorithm to calculate the objective function. Weighting factors allow you to adjust the significance of a particular criterion. For the synthesis problem for passive and active microwave circuits, the criteria can be ranked as follows: total length of connections, number of interlayer junctions, total resistance

of conductors and interlayer capacitance. The sum of all weights must be equal to one.

The designed hybrid genetic algorithm is a parallel algorithm based on a genetic algorithm with modified genetic operators. The structure of the genetic algorithm is similar to simple genetic algorithm excerpt for the use of modified genetic operators and consists of the following steps:

- Initialization, or selection of the initial population of chromosomes;
- Assessment of the chromosome fitness in a population;
- Testing the stopping conditions of the algorithm;
- Chromosome selection;
- Use of genetic operators;
- Formation of a new population;
- Selection of the 'best' chromosome.

The structure of hybrid genetic algorithm is depicted in Fig.5.

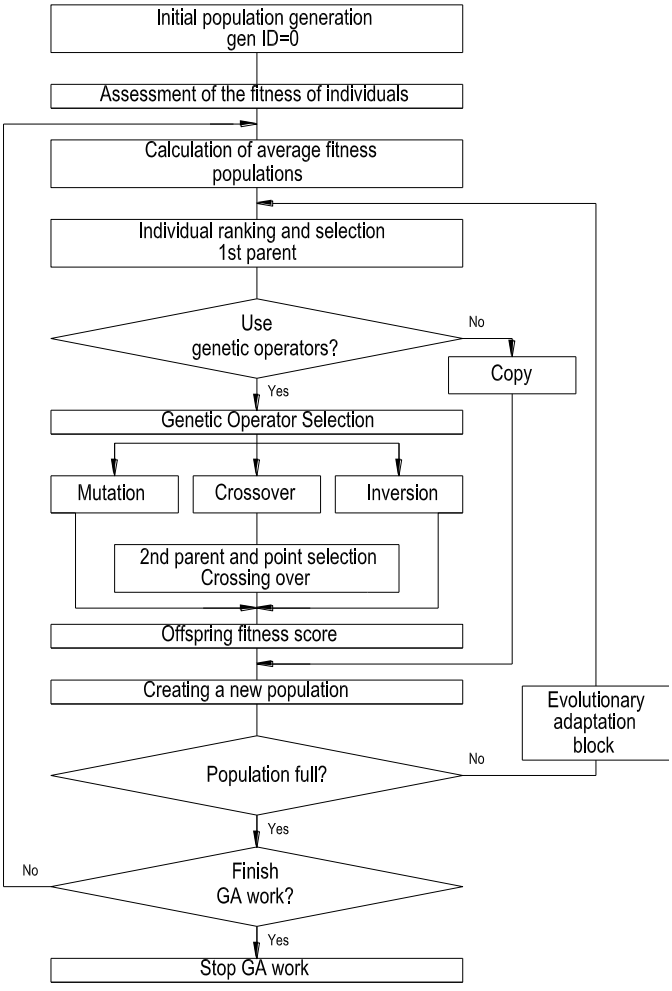


Fig. 5. Structural scheme of hybrid genetic algorithm

To evaluate the proposed solution, the algorithm was run for the difficulty and augmented dense switching blocks [14-16]. The values of the criteria for evaluating the algorithm are presented in table 1.

TABLE I. VALUES OF THE CRITERIA FOR EVALUATING

Test name	Algorithm	Number of transitions	Total length of connections, µm
difficulty	Weaver	65	323,25
	Montreal	68	315,21
	Proposed hybrid GA	63	283,58
augmented_dense	Weave	50	291,47
	Montreal	45	262,32
	Proposed hybrid GA	43	259,86

Despite the fact that the experiments were carried out under the same unchangeable conditions (operator parameters, number of iterations, population size), the time to solve the problem obtained with the same number of circuits and the number of leads is very far from each other. This is due to the varying complexity of the very sketch of the topology of passive and active microwave modules [10].

After the completion of the algorithm, the best determined solution is displayed on the screen in the form of a circuit topology of the microwave module. Table 2 shows the result of microwave amplifier calculation with the use of proposed methods.

TABLE II. EXAMPLE OF MICROWAVE AMPLIFIER CALCULATION

	N=3	**K1=1**	**K2=-1**	**K3=-1**	**K7=1**	**K8=1**	**K9=-1**		
	VA=50.0	VB=50.0	E=9.8	H=1.0	BK1=0.5	BK2=0.5			
	YA=2.0	WA=1.0	YB=2.0	WB=1.0					
	YC	DY	YT	WC	DW	WT	V		
1	12.000	2.0000	14.00	1.200	0.4000	1.6000	37.74		
2	12.000	2.0000	8.000	1.200	0.4000	0.8000	54.26		
3	12.000	2.0000	14.00	1.200	0.4000	2.0000	32.99		
4	20.000	4.0000	16.00	1.000	0.4000	1.4000	40.74		
5	20.000	4.0000	28.00	1.000	0.4000	1.0000	48.71		
6	20.000	4.0000	16.00	1.000	0.4000	1.0000	48.71		
7	12.000	2.0000	8.000	6.000	1.0000	8.0000	11.95		
8	12.000	2.0000	8.000	3.000	1.0000	1.0000	48.71		
9	12.000	2.0000	8.000	1.000	0.4000	1.8000	35.19		
10	20.000	4.0000	28.00	1.000	0.4000	1.8000	35.19		
11	20.000	4.0000	20.00	1.000	0.4000	0.2000	90.53		
12	20.000	4.0000	12.00	1.0000	0.4000	1.0000	48.7160		
	F	**RA**	**XA**	**RB**	**XB**	**VMA**	**RKCA**	**VMB**	**RKCB**
1	0.550	10.000	-5.00	2.000	6.000	0.144	1.337	0.51	3.13
2	0.600	11.000	-6.00	2.500	7.000	0.036	1.074	0.26	1.70
3	0.650	12.000	-7.00	3.000	8.000	0.076	1.165	0.35	2.10
	F	G1	G2	G3	GKPH	GKP	GKPD		
1	0.550	0.979	6.000	0.734	5.000	4.310	6.344		
2	0.600	0.999	5.500	0.932	5.000	5.122	7.094		
3	0.650	0.994	5.000	0.874	5.000	4.344	6.378		
	SU=0.02839			TU=0.69129					

For experimental studies, a hybrid genetic algorithm program was developed. The program is written in Borland® C ++ for operating systems of the Windows 98, NT, 2000, XP family. Experiments conducted on an IBM® compatible computer with an Intel® Xeon processor E5-2690, 20 MB cache, 2.90 GHz Turbo Boost 3.80 GHz (for deploying large populations in GA), 8.00 GT / s Intel® QPI Number of cores - 8, Number of threads - 16, server RAM 32GB. The program is registered in the register of computer programs by a certificate [17].

In addition to the identifiers indicated earlier and in the figures, the following are also used in the program: F – current frequency, in Hz; Assessment of the chromosome fitness in a population; N≤11 – the number of frequencies within the required range; VA, VB – wave impedance of the lines connected to the output and input of the amplifier, Ohm; BK1, BK2 – weighting factors V_1 and V_2 in the objective function; VMA, RKCA – reflection coefficient module and SWR from the output side of the amplifier, defined in relation to VA; VMB, RKCB –reflection coefficient module and SWR from the input side of the amplifier, defined in relation to the value VB; G1 and G3 – power transmission factors; G2 – transistor gain $K_{\text{P тр}}$; GKP – nominal gain of the entire amplifier; GKPD =101g GKP – the same gain, in decibels; GKPH – required simplifier gain.

The modified scheme of the genetic algorithm was presented and evaluated while calculating the transistor amplifier, which proved the efficiency of the scheme for solving problems of getting into local wells and premature convergence [19]. This calculation model allows effective problem solutions on multi-core processors.

IV. Conclusion

The article presents the bio-inspired approach to microwave circuit design based on GA, which makes it possible to obtain simultaneously a practically feasible circuitry solution providing for the specific features of manufacturing technology and initial version of the topology. The article also describes software calculations for the given approach implementation. The presented calculation example of a microwave amplifier confirms its effectiveness.

Acknowledgment

The reported study was funded by RFBR, project number 18-0700050.

References

[1] Danilchenko V.I., Kureichik V.M. Genetic algorithm for planning the placement of VLSI // Izvestiya SFU, vol. 2., 2019, pp. 75-79

[2] Lebedev B.K., Lebedev V.B. Planning based on swarm intelligence and genetic evolution // Izvestia SFedU. Technical science. - 2009. - No. 4 (93) - S. 25-33.

[3] Ahmet Karli,Vasfi Emre Omurlu, Utku Buyuksahin, Remzi Artar, Ender Ortak. Self tuning fuzzy PD application on TI TMS320F 28335 for an experimental stationary quadrotor. – URL: http://ieeexplore.ieee.org/document/6151404/ (date of the application 23.04.2017).

[4] A.A. Kalentiev, D.V. Garais, I.M. Dobush, L.I. Babak Structural and parametric synthesis of microwave transistor amplifiers based on a genetic algorithm using models of monolithic elements // TUSUR Reports, No. 2 (26), Part 2, December 2012, pp. 104-112.

[5] Tang, Maolin and Yao, Xin. A memetic algorithm for VLSI floorplanning // IEEE Transactions On Systems, Man, And Cybernetics–Part B: Cybernetics. – 2007. – № 37 (1).

[6] Goryainov A.E. Construction of parametric models of passive components of microwave monolithic integrated circuits using the Extraction-P program / A.E. Goryainov, I.M. Dobush, L.I. Babak // Nast. Sat. Pp. 94–99.

[7] Kokolov A. A., Salnikov A. S., Sheyerman F. I. and Babak L. I. Broadband Double-Balanced SiGe BiCMOS Mixer With Integrated Asymmetric MBaluns, Int. Conf. "Dynamics of Systems, Mechanisms and Machines" (Dynamics-2017), Omsk, Russia, 2017 (accepted for publication).

[8] Wenyuan L. and Qian Z. "A 0.7–1.9GHz Broadband Pseudo-diff erential Power Amplifi er Using 0.13-um SiGe HBT Technology". 2012 Int. Conf. on Microwave and Millimeter Wave Technology (ICMMT), pp. 1–4, July 2012.

[9] A. A. Kokolov, I. M. Dobush, F. I. Sheerman, L. I. Babak, et al. Complex functional blocks of broadband radio frequency amplifiers for single-chip L- and S-band receivers based on SiGe technology. 3rd Intern. scientific. conf. "Electronic Components and Electronic Modules" (International Forum "Microelectronics-2017"), Alushta, October 2017. - M .: Technosphere, 2017. - P. 395-401.

[10] Bocklemann D. E. and Eisenstadt W. R. Combined Diff erential and Common-Mode Scattering Parameters: Theory and Simulation. IEEE Trans. on Microwave Theory and Techniques, Vol. MTT-43, No. 7, pp. 520–523, July 1995.

[11] Kurokawa K. Power Waves and the Scattering Matrix, IEEE Trans. on Microwave Theory and Techniques, Vol. MTT-13, № 2, pp. 194–202, 1965.

[12] Zhabin D. A., Garays D. V., Kalentyev A. A., Dobush I. M. and Babak L. I. Automated Synthesis of Low Noise Amplifi ers Using S-parameter Sets of Passive Elements, Asia-Pacifi c Microwave Conference (APMC 2017), Kuala Lumpur, Malaysia, 2017 (accepted for publication).

[13] Kalentyev A. A., Garays D. V. and Babak L. I. Genetic-Algorithm-Based Synthesis of Low-Noise Amplifi ers with Automatic Selection of Active Elements and DC Biases, European Microwave Week 2014, Rome, Italy, pp. 520–523, October 2014.

[14] Babak L. I., Kokolov A. A. and Kalentyev A. A. A New Genetic-Algorithm-Based Technique for Low Noise Amplifier Synthesis, European Microwave Week 2012, Amsterdam, The Netherlands, pp. 520–523, November 2012.

[15] Mann G.K.I., Gosine R.G. Three-dimensional min–max-gravity based fuzzy PID inference analysis and tuning // Fuzzy Sets and Systems. – 2005. – Vol. 156. – P. 300-323.

[16] Kureichik V.M. Hybrid genetic algorithms // Izvestia SFedU. Technical science. - 2007. - No. 2 (77). - C. 5-12.

[17] Certificate 2020610223. "Software implementation of the hybrid algorithm for the placement of VLSI elements using a modified genetic algorithm": computer program / V. I. Danilchenko (RU), E. V. Danilchenko (RU), V.M. Kureichik (RU) .: copyright holder of the Southern Federal University. No. 2020610223; declared 01/09/2020; publ. 21.01.2020.

[18] A. A. Kokolov, I. M. Dobush, F. I. Sheerman, L. I. Babak, et al. Complex functional blocks of broadband radio frequency amplifiers for single-chip L- and S-band receivers based on SiGe technology. 3rd Intern. scientific. conf. "Electronic Components and Electronic Modules" (International Forum "Microelectronics-2017"), Alushta, October 2017. - M .: Technosphere, 2017. - P. 395-401.

[19] Zaporozhets D.Yu., Kravchenko Yu.A., Lezhebokov A.A., Methods of data mining in complex systems // Bulletin of the Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences. - 2013. - No. 3. - S. 52-54.

[20] Zhiqiang Yang, Jimin Zhang, Zhongchao Chen, Baoan Zhang. Semi-active control of high-speed trains based on fuzzy PID control // Procedia Engineering. – 2011. – Vol. 15. – P. 521-525.

Automatic Identification of Appendiceal Orifice on Colonoscopy Images Using Deep Neural Network

Anton Lebedev
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
lebedevdes@gmail.com

Vladimir Khryashchev
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
v.khryashchev@uniyar.ac.ru

Evgeniya Kazina
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
kazinaevgeniya@gmail.com

Anastasia Zhuravleva
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
an.zhuravleva76@yandex.ru

Sergey Kashin
Yaroslavl Regional Oncology Hospital
Yaroslavl, Russia
s_kashin@mail.ru

Dmitry Zavyalov
Yaroslavl Regional Oncology Hospital
Yaroslavl, Russia
zavialoff@mail.ru

***Abstract*— The results of testing the recognition algorithm of the cecum achievement in colonoscopy video of the colon mucosa are presented. The image database was formed from the results of colonoscopy procedure together with the doctors of the Yaroslavl Regional Oncology Hospital. As the architecture of the convolutional neural network, the ResNet50 modification, previously trained on the standard ImageNet base, was chosen. As a result of applying the machine learning algorithm to the test set of endoscopic images, the metric values were AUC=0.95, F-score=0.9, when a threshold is h=0.462. The results can be used to develop a quality control system for colonoscopy procedures. The introduction of such a system in medical practice will partially automate the analysis of video data, which will subsequently lead to a decrease in the number of subjective medical errors during colonoscopy.**

Keywords*— *Colonoscopy image analysis; convolutional neural network; deep learning; computer vision

I. INTRODUCTION

Today, computer vision systems in combination with artificial intelligence methods are used in different areas of life. One of the directions in the development of such systems is the use of machine learning algorithms to solve the problems of automatic detection and classification of target objects in images. Progress in this area and the development of appropriate software and hardware technologies for computer vision makes it more realistic to create automatic diagnostic systems, as well as decision support systems [1, 5, 12]. The introduction of such systems in clinical medicine is aimed to increasing the efficiency of diagnostics and therapy, reducing the time and cost of research, conducting quality control, as well as training and improving the medical skills of specialists.

With the help of deep learning, the tasks of classification, segmentation and detection that arise in clinical practice in the medical images analysis represented by various visualization methods (radiography, computed tomography, MRI) are actively being solved [4]. At the moment, it is known about the use of deep learning in areas such as detecting breast cancer on mammograms, segmentation of liver metastases using computed tomography (CT), segmentation of a brain tumor using magnetic resonance imaging (MR), classification of interstitial lung diseases with using high-resolution chest CT and the creation of appropriate tags related to the contents of medical images.

In particular, one of the current research areas is the analysis of endoscopic images [4-5]. According to estimates of the World Health Organization by the middle of the XXI century diseases of the gastrointestinal tract will occupy one of the leading places, which is largely due to the ecology and lifestyle of modern man. In addition to taking preventive measures, an important role in the fight against gastrointestinal diseases is played by their early diagnosis, carried out, in particular, by conducting endoscopic examination. During endoscopy, an endoscope is inserted into the cavity of the human organ, which allows the mucous membrane and the internal structure of the digestive tract to be displayed on the screen, including using various operating modes, such as increasing the studied area of the mucous membrane, using a narrow spectrum of light, using various dyes, etc.

Visually on endoscopic images, this pathology has its own distinctive features, which allows the doctor to attribute it to a benign or malignant formation. However, the specialist's level of training, his physical and emotional state, as well as fatigue can have a significant impact on the quality of diagnosis and the likelihood of medical errors.

Endoscopic research is central to the diagnosis of stomach cancer and colon cancer, and the use of additional endoscopic technologies increases the efficiency of precancerous pathology detection and early forms of cancer [6]. Today there are a number of studies on systems for automatic endoscopic images analysis.

The article [7] presents a system for detecting gastric cancer in the early stages, based on a convolutional neural network. To test the network, authors used data collected after 68 endoscopic

The reported study was funded by Russian Foundation for Basic Research (RFBR), project number 19-37-90153.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

examinations of the stomach, 62 of which confirmed the presence of cancer. The accuracy of this network was 94.1%, and the average time for pathology detection was 1 second.

The system of automatic classification of tumors on the stomach walls is described in [8]. It was based on a convolutional neural network. The weighted average accuracy reached 84.6% for classification into five categories. The average area under the curve (AUC) of the model for the differentiation of gastric cancer is 0.877, for the differentiation of neoplasms – 0.927.

In [9], the authors proposed a solution for the polyps automatic detection on the gastric mucosa using a convolutional neural network. According to the results, this development operates in real time at a speed of 50 frames per second (FPS) and guarantees an average accuracy (mAP) from 88.5% to 90.4%.

The above analysis of the literature shows that the use of computer vision systems for the endoscopic images analysis gives good results today, according to numerous medical studies. Thus, the development of such systems is an urgent scientific and technical task.

Despite the absence of a uniform standard for colonoscopy in the world it is possible to assess the quality of the procedure by observing a number of requirements by a specialist. So, the completeness of the examination is evidenced by bringing the endoscope to the patient's cecum, where the colonoscopist performs photo fixation of appendiceal orifice [1].

Endoscopic analysis is subjective as the specialist needs to determine and fix points of interest independently. At the same time, due to structural features of the colon (large organ length, heterogeneity, gaps, the presence of several anatomical sections, folds), colonoscopy is considered a complex procedure even for professional endoscopists. These factors increase the risk of medical errors [1, 4].

A number of studies focus on the use of deep learning in colonoscopy. In paper [2] a classification system for video frames of a colonoscopy examination is presented, based on a convolutional neural network with binarized scales. The Dice coefficient for the proposed solution is 71.20%, and the accuracy exceeds 90%. The authors of [3] proposed a model based on deep learning, capable of differentiating adenomatous and hyperplastic colorectal polyps in real time. The detection accuracy of polyps was 94%, the sensitivity of identification by adenomas was 98%, the specificity was 83%. The convolutional neural network developed in [10] also copes with the task of detecting and classifying polyps. With its help, the accuracy of diagnostics carried out by specialists increases to 85.9%. In [11], the authors managed to achieve an accuracy of colon sections classification of about 90% through the use of a convolutional neural network with deeply trained hierarchical features.

This article has the following structure. The second part describes the base of colon endoscopic images used to train and test the developed algorithm. The third part presents a solution for appendiceal orifice recognizing on colonoscopy images, which was based on the ResNet50 convolutional neural network architecture. The fourth part describes the training and testing of the algorithm, as well as the numerical values of the test results.

II. Colonoscopy Images Database Description

In this work a database compiled by specialists of the Yaroslavl Regional Oncology Hospital was used. Images were received during a colonoscopy by endoscopic photo fixing of different parts of the patients colons, including the appendiceal orifice. The database was represented by images with resolutions of 624x528 pixels and 640x480 pixels, obtained from the following endoscopic systems: endoscopic system OLYMPUS EXERA II, video gastroscope GIF 160Z; endoscopic system OLYMPUS LUCERA SPECTRUM, video gastroscope GIF 260Z; endoscopic system OLYMPUS EXERA III, video gastroscope GIF HQ290; endoscopic system PENTAX Medical EPK-i7010, video gastroscope EG-2990Zi.

From an algorithmic point of view, the binary classification problem was solved, where the first class represented images of the colon mucosa, the second class - images with the appendiceal orifice (Fig. 1).

Distinctive features of the appendiceal orifice are the following physiological signs:

- a curved hole forming at least one quarter of a circle or oval with bounding folds;

- a closed hole with two or more surrounding circular or almost round folds;

- wide open hole without adjacent surrounding folds.

The image database was collected from the results of colonoscopy procedure together with the doctors of the Yaroslavl Regional Oncology Hospital. At the moment, this image database consist of 2671 images. Among them 2294 is a negative class (without the appendiceal orifice), 377 positive (containing the appendiceal orifice). This base was randomly divided in the ratio of 80% to 20% into a training and validation set. Thus, the training base consists of 2136 images, of which 311 are positive and 1825 are negative. The validation set consists of 535 images, including 66 positive images and 469 negative ones. In addition, doctors manually collected a test data set of 104 images, of which 57 were positive and 47 were negative. This database contains complex cases - images of diverticula, which are visually similar to the orifice of the cecum of a diverticular type, various types of the orifice of the cecum images and images of pathologies.

III. Neural Network Architecture

The convolutional neural network ResNet50 was used for classification appendiceal orifice. This neural network was pre-trained on an ImageNet dataset. The ResNet50 architecture showed a high result in modern problems of object detection and classification on digital and medical images, at the same time, the computational complexity of this architecture allows it to be used for processing a video stream in real time, which is important for developing a colonoscopy quality control system based on of this algorithm. The neural network architecture was modified by replacing the output layer with two fully connected layers with 1024 neurons. A dropout with a probability of 0.5 in fully connected layers was used for the regularization of the model during training. The output layer is binary classifier with sigmoid as function activation. Table 1 represented a structure of the trained neural network.

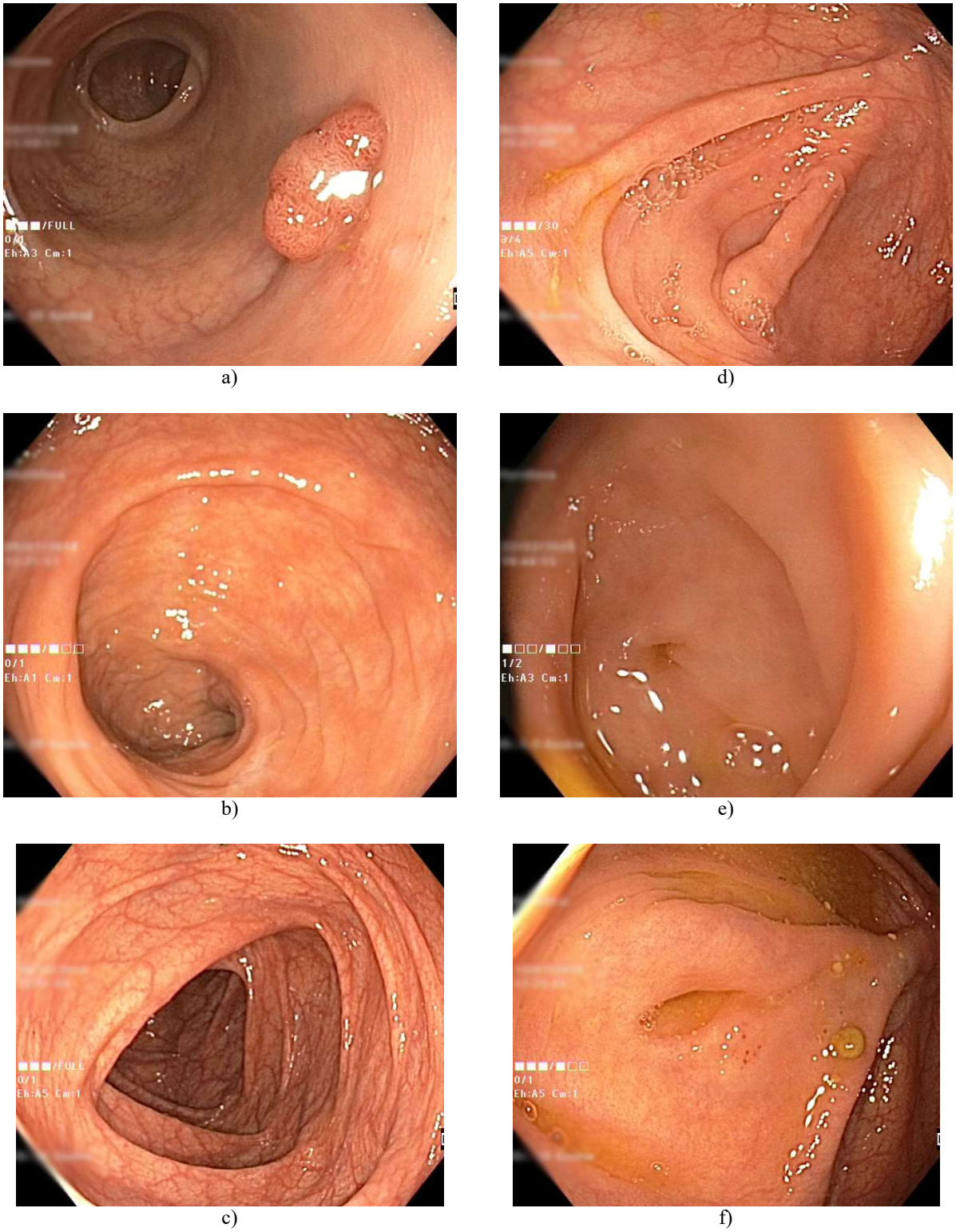


Fig. 1. Examples from the colonoscopy images database: a, b, c - colon mucosa; d, e, f - the appendiceal orifice.

TABLE 1. THE ARCHITECTURE OF THE USED RESNET50 NEURAL NETWORK

Layer name	Output_ size	Layer/Block Layers	Replay
Conv1	112×112	$7 \times 7, 64, stride\ 2$	1
MaxPool2D	56×56	$3 \times 3 pool, stride\ 2$	1
Conv2_x	56×56	$[1 \times 1, 64\ 3 \times 3, 64\ 1 \times 1, 256\ ]$	3
Conv3_x	28×28	$[1 \times 1, 128\ 3 \times 3, 128\ 1 \times 1, 512\ ]$	4
Conv4_x	14×14	$[1 \times 1, 256\ 3 \times 3, 256\ 1 \times 1, 1024\ ]$	6
Conv5_x	7×7	$[1 \times 1, 512\ 3 \times 3, 512\ 1 \times 1, 2048\ ]$	3
Gl_average_pool	1×1	*average pool layer*	1
FC1	1024	*fully connected layer*	1
FC2	1024	*fully connected layer*	1
Sigm	2	*sigmoid function layer*	

IV. NEURAL NETWORK TRAINING AND TESTING RESULTS

For training network we used the training dataset contained 2671 images described in detail in paragraph 2, and we used validation dataset for evaluate model at the end of each training epoch. During the training we used set of random transforms such as the horizontal flip, vertical flip, rotation in range 20 degrees for data augmentation.

For layers from the base part of the network, pre-trained on the ImageNet dataset weights were used for initial initialization. Additional layers was initialized by Xavier initializer.

Since the classes are not balanced, for better convergence of the initial bias of the output layer was set as:

$$\text{bias} = log\ \left(\frac{N_{pos\_nb}}{N_{neg\_nb}}\right),$$

also, for weighting the loss function, were added weight values for each class calculated by the formulas below:

$$w_{neg\_class} = \frac{N_{total}}{2 * N_{neg\_nb}},$$

$$w_{pos\_class} = \frac{N_{total}}{2 * N_{pos\_nb}}$$

where N_{total} – number of total samples, N_{pos_nb} – number of positive samples, N_{neg_nb} – number of negative samples, w_{neg_class} – weight for the negative class, w_{pos_class} – weight for the positive class.

The neural network has been trained for 90 epochs. The Adam (adaptive amount estimation) with a learning rate of 0.001 and decay of lr=1e-3 was used as the optimizer. The binary cross-entropy was used as the loss function.

The ROC-curve was used to analyze the algorithm. In Fig. 2, the ROC-curve for the validation set of endoscopic images is presented. The best result on the validation set was AUC = 0.97. Additionally, the F_1-score metric was used in the evaluation of the algorithm. The best value is F_1-score = 0.85 in the validation dataset, when a threshold is th = 0.608

Then the trained model was checked on a test set, described in detail in paragraph 2. In Fig. 3, the ROC-curve for the test set of endoscopic images is presented. In this case, the area under the curve is equal to AUC = 0.95.

The F_1-score for the test set was computed for threshold which was found on validation set th = 0.608 and equal 0.89. In the same time the greatest F_1-score value for test set is equal 0.9 with a threshold th = 0.462. The average analysis time of one image is 29 ms, which allows to process up to 40 images per second. Performance specs were obtained on an Nvidia GTX 980Ti GPU, without optimizations or TensorRT acceleration. These results indicate that the model has a good generalizing ability. The research showed that the trained network had a high F_1-score value on the test set. These results can be used in the development of a colonoscopy quality control system.

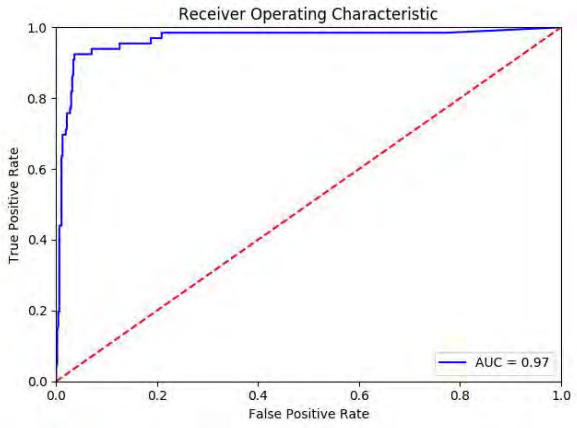


Fig. 2. The ROC-curve for the validation set of colonoscopy images.

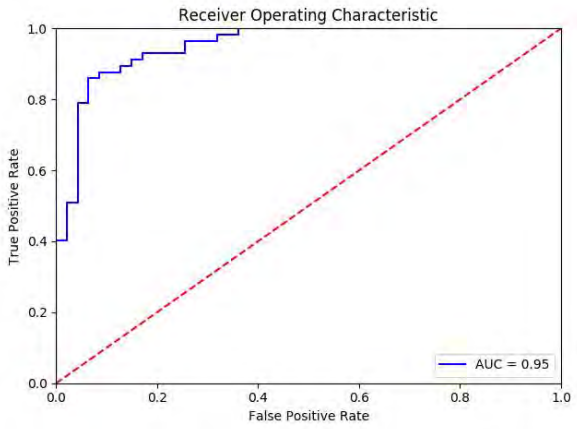


Fig. 3. The ROC-curve for the test set of colonoscopy images.

V. Conclusion

The classification algorithm is proposed and tested for the appendiceal orifice in a caecum area. The convolutional neural network based on the ResNet50 architecture was trained and testing.

The following results were obtained on a test dataset in the process of the study – AUC=0.95, F-score=0.89, when a threshold is th=0.608. This score is a high result for object classification task in the endoscopic images. The researchers plan to collect a more representative database of colonoscopy images collaboratively with specialists from the Yaroslavl regional cancer hospital to improve the performance and generalizing ability of the trained neural network.

These results can be used in the development of a quality control system for conducting colonoscopy procedure. The introduction of such a system in medical practice will partially automate the analysis of video data. These will lead to a decrease in the number of subjective medical mistakes during colonoscopy.

Acknowledgment

The research was carried in collaboration with doctors from the endoscopic department of the Yaroslavl Regional Oncology Hospital. The authors are also grateful to the Center for Artificial Intelligence of Yaroslavl State University for providing access to the NVIDIA DGX-1 supercomputer.

References

[1] Münzer, B., Schoeffmann, K., and Böszörmenyi, L. 2018. Content-based processing and analysis of endoscopic images and videos. A survey. In: Multimedia Tools and Applications. 77, 1323–1362.

[2] Byrne, M., Chapados, N., Soudan, F., Oertel, C., Linares Pérez, M., Kelly, R., Iqbal, N., Chandelier, F., and K Rex, D. 2017. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. In: Gut (2019 Jan). 68(1), 94-100. DOI=10.1136/gutjnl-2017-314547. Epub 2017 Oct 24.

[3] Zheng, Y., Zhang, R., Yu, R., Jiang Y., Mak T.W.C., Wong, SH., Lau, J.Y.W., and Poon, C.C.Y. 2018. Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features From Nonmedical Domain. In: Conf Proc IEEE Eng Med Biol Soc (2018 Jul). 4142-4145. DOI=10.1109/EMBC.2018.8513337.

[4] Zhou, S.K., Greenspan, H., and Shen D. 2017. Deep Learning for Medical Image Analysis. Elsevier Science.

[5] Khryashchev, V., Lebedev, A., Stepanova, O., and Srednyakova, A. 2019. Using Convolutional Neural Networks in the Problem of Cell Nuclei Segmentation on Histological Images. In: Recent Research in Control Engineering and Decision Making (ICIT 2019). Studies in Systems, Decision and Control. Vol 199. Springer, Cham.

[6] Kuvaev, R.O., and Kashin, S.V. 2016. Modern endoscopic examination of the stomach using narrow-spectral and magnifying endoscopy techniques: techniques and diagnostic algorithms. In: Evidence-based gastroenterology. 2 (5), 3-13.

[7] Mitsuaki, I., Toshiaki, H., and Tomohiro, T. 2018. Detecting gastric cancer from video images using convolutional neural networks. In: DEN Video Article (18 November 2018). DOI=https://doi.org/10.1111/den.13306.

[8] Bum-Joo, Ch., Chang, S.B., Se, W.P., Young, J.Y., Seung, I.S., Hyun, L., Woon, G.S., J, T.H, Yong, T.Y., Seok, H.H., Jae, H.C., Jae, J.L.,and Gwan, H.B. 2019. Time for second-generation artificial intelligence in medical imaging. In: Endoscopy. 51(12), 1113-1114. DOI=10.1055/a-0999-5476.

[9] Zhang, X, Chen, F, Yu, T, An, J, Huang, Z, and Liu, J. 2019. Real-time gastric polyp detection using convolutional neural networks. In: PLoS ONE. 14 (3), e0214133. DOI=https://doi.org/10.1371/journal.

[10] Park, S., Lee, M., and Kwak, N. 2015. Polyp detection in Colonoscopy Videos Using Deeply-Learned Hierarchical Features, Technical Report. Seoul National University.

[11] Heresbach, D., Barrioz, T., Lapalus, M., Coumaros, D., Bauret, P., Potier, P., Sautereau, D., Boustiere, C., Grimaud, and J., Barthelemy, C. 2008. Miss rate for colorectal neoplastic polyps: a prospective multicenter study of backto-back video colonoscopies. In: Endoscopy. 40, 4, 284–290.

[12] Khryashchev V., Stepanova O., Lebedev A., Kashin S., Kuvaev R. 2019. Deep Learning for Gastric Pathology Detection in Endoscopic Images. ACM International Conference Proceeding Series, 3rd International Conference on Graphics and Signal Processing, ICGSP, 90-94.

Automatic control system for dedusting of gas-cleaning plant filtering element

Anton Zyma
Department of Systems Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
zima.ae@gmail.com

Leonid Rebezyuk
Department of System Engineering
Kharkiv Naional University of Radio Electronics
Kharkiv, Ukraine
https://orcid.org/0000-0001-8516-6584

***Abstract*— Filtration of dusty air obtained in the production process is today the most important and necessary task for any industrial enterprise. To maintain the efficiency of the filter elements, they are cleaned with compressed air (dedusting). The adaptive choice of the start and end points of time of filter purging determines the energy efficiency and productivity of the entire cleaning system. With a view to increase them, a multifunctional regeneration control device was developed.**

Keywords—filter, gas-cleaning plant, automatic control system, microcontroller, differential pressure sensor, adaptative control

I. INTRODUCTION

Along with the transition to environmentally friendly energy resources in all industry branches, their optimal and efficient use, without losing the quality of the tasks performed, today occupies a leading position in the world among the issues studied by science [1]. In particular, with the existing and widespread method of filtering dust and gas flows using high pressure filters, there is a problem of cleaning them with inevitable contamination over time. Nowadays, filter dedusting control systems perform regeneration at constant time intervals. A method for determining these time intervals at the stage of pre-design work is proposed in [2]. The main goal of this work is to develop a model for adaptive control of the filter dedusting of a gas-cleaning plant based on the hydraulic resistance of the filter, as well as the development of a control system for the filter elements cleaning process, which provides adaptive time intervals definition and additional functionality for controlling third-party equipment, displaying and setting the values of the controlled parameter (including from a mobile device, for example, a smartphone or tablet PC), implementing M2M (Machine-to-machine) communication via a wireless protocol (Bluetooth 5.0, Thread or ZigBee). The combination of characteristics and versatility of the developed controller makes it possible to expand the scope of its application without significant design changes.

II. FILTRATION PROCESS

This section briefly describes the common operating phases of the gas-cleaning plant filtering element (filter sleeve).

All work on separating particles from gas is divided into two phases (Fig. 1): Filtration of dust-loaded gases and cleaning of the filter sleeves.

A. Filtration of dust-loaded gases

The dust-containing gas is penetrating from the untreated gas chamber (Fig. 2 (2)) through the filter sleeves (Fig. 1 (1)).

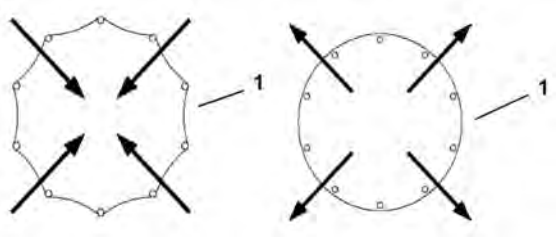
The dust particles settle on the surface of the filter sleeves and the cleaned gas escapes in an axial motion through the filter sleeves. To prevent the filter sleeves (Fig. 1 (1)) from getting pressed together, there is a sleeve cage inside. The diameter of the filter sleeve is a little bigger than that of the sleeve cage so that the filter sleeve lies in the form of a garland around the lengthways rods of the sleeve cage in normal operation. The cleaned air escapes at the top end of the filter sleeve and is collected in the clean gas chamber (Fig. 2 (3)). From there the cleaned air passes to the fan at the filter outlet (Fig. 2(8)).

This phase is common to this kind of separator and cannot be changed.

B. Cleaning of the filter sleeves

In the second phase (Fig. 1b) the filter bags are cleaned using a purge air counterblast of 5 bar from the purge air nozzle, which is directly built into the purge air tank. The time interval of the purge air blast is set on the filter control. The purge air blasts are made continuously. Each filter sleeve is individually dedusted.

The filter sleeve (Fig. 1(1)) is inflated abruptly and the layer of filter dust that sticks on the outside is broken up and thrown off. The dust falls down into the silo or into the outlet cone (total separator or aspiration cone).



(a) (b)

Fig. 1. Filter sleeve operating phases (1): (a) Filtration of dust-loaded gases; (b) Cleaning of the filter sleeve

There are several different filter dedusting approaches, for instance, mechanical cleaning, which assumes that filter bags

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

are stressed through shaking during the cleaning operation. However, the shown method with periodic reversal of the flow

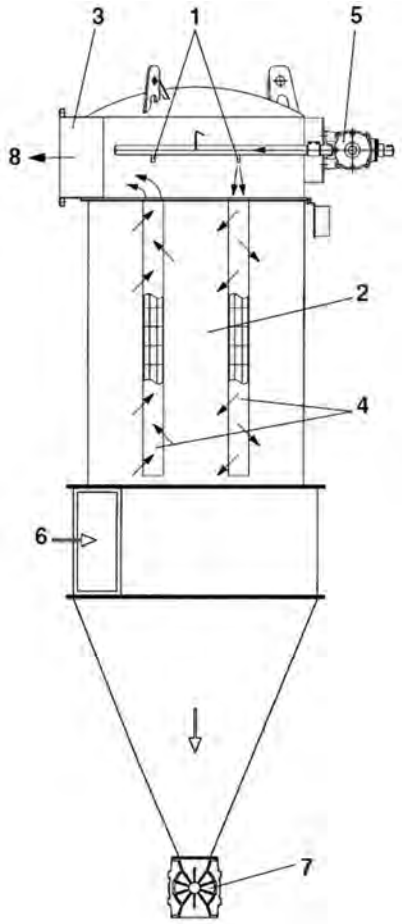


Fig. 2. High pressure filter: (1) Purge air nozzle; (2) Untreated gas chamber; (3) Clean gas chamber; (4) Filter sleeves; (5) Purge air tank; (6) Aspiration inlet; (7) Dust discharge gate; (8) Clean gas outlet

direction (reverse flow filter) is a much gentler [3]. The filter system here features several separate chambers which are cleaned individually.

So, for example, similar gas cleaning plants are used in feed mills, where air, which is polluted at all stages of production with fine- or coarse-dispersed dust in aerosol form is filtered. As a rule, the filter sleeves of the treatment plant are made of polyester.

At present, the dedusting periods and the intervals between them in the treatment plants are constant, i.e. according to time-relay actuation and are not adaptive. Of interest is an approach that provides for cleaning the filter sleeves when a critical pressure difference between the chambers of the cleaned and untreated (dusty) gas is reached, which allows the control system to adaptively define the dedusting time intervals and thereby increase the efficiency and reduce the energy consumption of the gas cleaning plant.

III. Mathematical conditions for adaptation of filter dedusting parameters

To achieve the maximum efficiency of the filtration process, it is necessary to determine the values of the hydraulic resistance of the filter element at which the dedusting process starts and stops.

In this case, the first phase of the filtration process will be performed when the value of the filter hydraulic resistance changes in the range $[\Delta P_{min}, \Delta P_{max}]$.

Since a porous filter partition, in the general case, can be represented as consisting of two layers [4]: the primary, which is a porous partition proper with dust particles deposited on the walls of the pore channels and the secondary is a layer of trapped particles accumulated on the frontal surface of the filter, the total hydraulic resistance can be represented as (1).

$$\Delta P = \Delta P_1 + \Delta P_2, \tag{1}$$

ΔP_1 and ΔP_1 here are hydraulic resistance of the primary and secondary filter layers, respectively.

Determining the value of ΔP_{min} it is necessary to take (2) into account.

$$\lim_{\Delta P \to \Delta P_1} H_\Sigma = H_1, \tag{2}$$

H_Σ denotes total filter element thickness, H_1 – thickness of the filtering partition primary layer.

Condition (2) illustrates that ideally, when total hydraulic resistance (differential pressure) of the whole filter surface tends to the resistance of the primary filter layer, we are losing the secondary layer. Thus, total filtration element thickness tends to the thickness of the primary layer.

Consequently, taking into account (2), as well as the data [4], we can assume that:

$$\Delta P_{min} > \Delta P_2. \tag{3}$$

If condition (3) is met, then, accordingly, the presence of a minimum amount of dust in the secondary layer is excluded, which will provide a higher filtration quality together with a limitation of energy consumption for filter dedusting by reducing the second sleeve operating phase duration. Also, when calculating, it is necessary to maximize the filtration rate included in its composition (4) to ensure the maximum performance of the gas-cleaning plant.

Based on [4]:

$$\Delta P_2 = (B\mu Z\omega^2)t, \tag{4}$$

B – dust layer resistance coefficient (m/kg), μ – dynamic gas viscosity coefficient (Pa*s), Z – dust content of the gas in front of the filter under operating conditions (kg/m^3), ω – filtering speed (m/s), t – time counted from the start of filter operation (s).

The value of B in (4) is determined empirically only.

When calculating the value of the upper limit of the differential pressure, it is noteworthy that its value should be as high as possible, since this will increase the periods between dedusting cycles and reduce energy costs accordingly. In this case, the filtering element must provide the required filtering rate (5), thereby ensuring the purification of the entire volume of gas to be purified.

$$\omega_f \geq \omega_{fn}, \tag{5}$$

ω_f denotes filtering speed (m/s), ω_{fn} – required filtering rate (m/s).

From the above it is clear that it is impossible to find the universal values of ΔP_{min} and ΔP_{max}. Their values depend on the nature of the filtered gas pollution (type of dust), the porous partition material and the required filtration rate. Therefore, the differential pressure values for dedusting should be calculated for a particular gas-cleaning plant when applying in a specific production.

IV. CONTROL SYSTEM

This section briefly describes the developed universal high-pressure filter dedusting control device.

Based on the above, an adaptive microcontroller control system with a relay control law has been developed to solve the problem. The block diagram of the system is shown in Fig. 3.

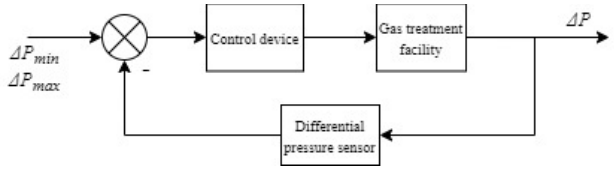


Fig. 3. Automatic control system block diagram

Among the functional requirements put forward to the system that is being developed, the following can be distinguished:

1) Sensors: In order to determine the current value of the differential pressure between the chambers of clean and untreated gas, it is necessary to use a differential pressure sensor.

2) Inputs and outputs: The control functions of this system are implemented through discrete outputs in the amount of 4 pieces. To support additional functionality in tasks of control and monitoring of third-party equipment, it is also necessary to provide 4 dry contact inputs, 1 analog input and 1 analog output. The voltage at the analog I/O should vary between 0-10 VDC. For digital inputs and outputs, the voltage can be either 0 VDC or 24 VDC.

3) Interfaces: It is necessary to realize three types of interfaces: human-machine (to set the controlled parameters and display the required information), wired (to implement the possibility of connecting this system to existing communication channels at the enterprise) and wireless (used to display information and change the values of the controlled parameter on the connected smartphone or tablet PC, as well as to enable the implementation of a mesh-network of similar devices in the enterprise, together with the microcontroller software update over the air).

4) Information displaying: A touch-sensitive color LCD-display must be used as an indication device. Such a display will make possible an intuitive user interface realization and ensure the information readability.

5) Supply voltage: 24 VDC.

6) Additional requirements: In order to guarantee safe and interference-free operation of the device, it is necessary to ensure complete galvanic isolation of all its inputs and outputs (including power lines) from internal systems (e.g. microcontroller, sensor, display). Also, to ensure the high-quality operation of the filter dedusting system under conditions of periodic power outages, it is important to provide the microcontroller dedicated periphery with uninterruptible power supply for up to several days. The system must guarantee the availability of a real time clock with a calendar to save some statistics of its own functioning.

The block diagram (Fig. 4) shows the main elements of the developed control system. We will briefly describe them below:

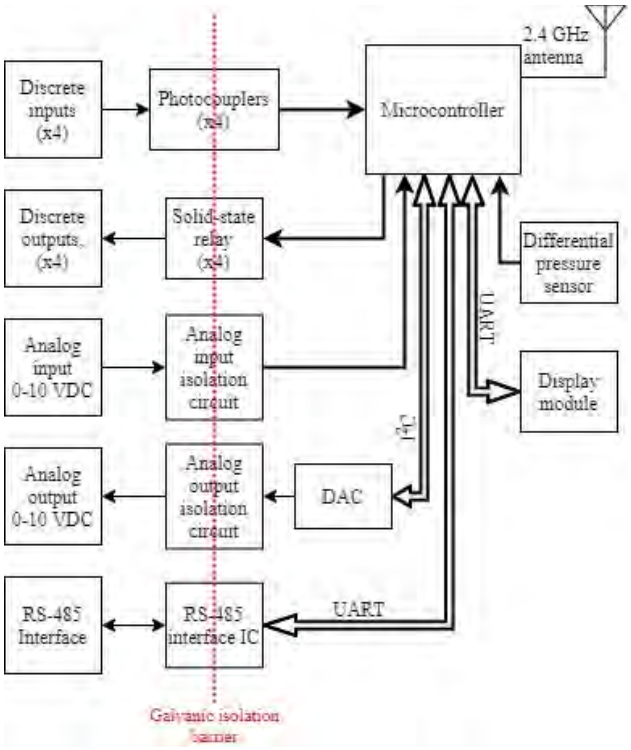


Fig. 4. Control device block diagram

a) Photocouplers are used to isolate discrete inputs of a device and a microcontroller. They also implement galvanic isolation and matching of signal levels at the input of the device and the microcontroller.

b) Solid-state relays act in the same way as photocouplers except that they implement discrete outputs. The use of solid-state relays is due to the low power of the loads that need to be controlled, as well as the greater wear resistance of such relays because of the absence of mechanical contact inside of them.

c) Analog input and output isolation circuits required for galvanic isolation and matching of the input and output analog signal levels.

d) RS-485 interface integrated circuit acts as an interface converter (from RS-485 to UART), and also implements opto-isolation of signals.

e) DAC (Digital to analog converter) is used as a separate integrated circuit since the selected microcontroller does not have a built-in digital-to-analog converter.

f) Microcontroller selected from the point of view of the maximum coverage of all the device functionality necessary for the implementation of its own capabilities. This approach reduces costs on the outer periphery. The selected STM32WB series IC represents the latest line of microcontrollers from ST Microelectronics, the main feature of which is the presence of a separate ARM Cortex M0+ core specifically to operate with the network protocol.

g) Display module for displaying information and entering the control system reference parameters.

h) Differential pressure sensor with a pressure measuring range from -2 kPa to 2kPa.

i) 2.4GHz antenna is realized on the PCB according to the microcontroller manufacturer requirements.

So as to meet all additional requirements to the current system, the power line was equipped with the supercapacitor, which is used to power the RTC (real-time clock) and some backup registers of the microcontroller. The registers applied to keep statistics and the system's current state in case of power outage. The use of a supercapacitor instead of a battery avoids problems with its replacement in the future, which increases the reliability of the developed device.

The appearance of the developed device is shown in the fig. 5.

The algorithm for adaptive control of the gas cleaning plant filter dedusting process can be represented as follows:

1) Check backup register dedusting bit: if 1 then go to step 7;

2) Get the actual differential pressure (ΔP) value from the sensor;

3) Send ΔP value to the display module via UART;

4) Compare the actual ΔP value to the preset boundaries ΔP_{min} *and* ΔP_{max}, ΔP_{min_alarm} *and* ΔP_{max_alarm}: here ΔP_{min_alarm} and ΔP_{max_alarm} are the parameters denoting the minimum and maximum hydraulic pressure values, which are used to inform the operator about an emergency situation (e.g. filter sleeve is torn hence no hydraulic pressure at all);

5) If ΔP exceeds ΔP_{max} then go to step 6, otherwise go to step 8;

6) Set a bit to the backup registers with a timestamp: this action is done so as to continue an unfinished dedusting process, which could be interrupted by the power outage;

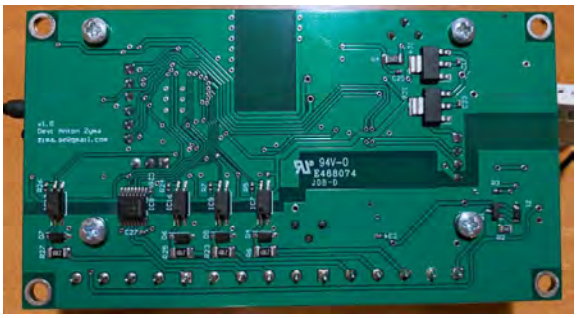
7) While ΔP exceeds ΔP_{min} and at least one discrete input is activated (24VDC on input), set one discrete output to high level;

8) Check for the command or data from the dispalay module via UART: when user sets the new ΔP boundaries, new date/time or sensor calibration command, the display module sends it together with the serial interface;

9) If new data arrived, overwrite non-volatile memory values: ΔP boundaries, date and time are stored in non-volatile microcontroller memory (EEPROM);

10) Go to step 2.

(a)



(b)

(c)

Fig. 5. Control device photos: (a) Top view; (b) Bottom view; (c) Top view with the display module installed

It should be noted that the algorithm represented above does not fully reflect the functioning of the system, because, along with the interrupts mechanisms or direct memory access, the real-time operating system (FreeRTOS) is used. It introduces some multitasking to the whole system, that leads to the simplified scalability and better device response.

V. CONCLUSIONS

Thus, in the filtration control process (dedusting), the system allows to adaptively determine the time intervals between regenerations and purge pulses in real time based on information about the pressure at the filter inlet and outlet. Pilot studies are planned to determine the efficiency of the regeneration system in terms of filter throughput and energy consumption for the filtration process.

REFERENCES

[1] Patel, R., 2010. Overview of industrial filtration technology and its applications. Indian Journal of Science and Technology, 3(10), pp.1121-1127.

[2] S. Zotov, "Improvement of the automated system control of dry gas cleaning plants of the aluminum plants", Ph.D., Ural Federal University, 2011. (in Russian)

[3] Schrooten, T., Kögel, A., Daniel, T. and Klein, G., 2010. Industrial dedusting with bag filters. Global guide of the filtration and separation industry, pp.156-160.

[4] Ladygichev, M. and Berner, G., 2004. Foreign And Domestic Gas Purification Equipment. Moscow: Teplotehnik.(in Russian)

Iterative Methods for Multi-Valued Logical Equation System Solving while Digital System Simulating

Alexander Ivannikov
Institute for Design Problems in Microelectronics of Russian Academy of Sciences
Moscow, Russian Federation
adi@ippm.ru

***Abstract*—The article is devoted to the analysis of methods for solving systems of multivalued logical equations by iteration methods. Iterative methods for solving such systems of equations are a mathematical description of the main process of functional-logical simulation, which is used at the design stage of digital systems to check the correctness of the design. Consideration of multivalued (finite-valued) values of logical signals at the outputs of blocks and elements of digital systems is explained by the fact that in a number of cases, to analyze the correctness of time relationships when simulating of digital system hardware, several-valued representations of binary logical signals are used, as well as those that recently, the development of logical elements that implement four or more significant logic is underway. Based on the analysis of the structure of the system of logical equations used in digital equipment simulation, using graph and logical models, an analysis of the existence of solutions and their number is carried out. Iterative methods of simple and generalized iteration are analyzed, the relationship between the number of solutions to a system of equations and its graph representation, reflecting a given connection scheme of elements of a digital system hardware, is shown. For the generalized iteration method, variants with different structures of the iteration trace are considered, in particular, it is shown that with a certain structure of the iteration trace, the generalized iteration turns into a simple iteration or Seidel iteration. It is shown that the generalized iteration most adequately describes the process of simulating the switching of logical signals in digital system hardware. The correspondence between various variants of functional-logical simulation of digital systems and the used method of iterative solution of systems of logical equations is shown.**

Keywords—functional-logical simulation of design, digital systems, multivalued logical equations, methods of simple and generalized iteration

I. Introduction

When designing digital object management systems, functional-logical modeling at the level of connecting blocks and circuit elements of hardware is necessary to verify the correctness of the design. To verify the correct functioning of the developing control system, simulation of changes in logical signals is carried out both at the output of the entire control system and at the nodes of the connection of blocks and elements of the hardware of the digital system. Moreover, for each change in the input signals, it is necessary to determine new logical values of the signals in the nodes of the circuit [1-3], that is, to implement a solution of a system of multi-valued (having course significant) logical equations. The ambiguity of logical variables is due to the presence, in addition to states 0 and 1, of a high impedance (disabled) state on the lines and buses of digital blocks, the representation of signals on the buses as a single multi-valued signal [4, 5], and also the representation of the process of switching signal values in the form of several significant logical signals, which used in modeling to verify time diagrams [6]. In addition, digital blocks are currently being developed that use a four-digit representation of logical signals, as well as logical signals of a different significance [7-10].

The specific form of the system of equations, as well as the fact that models of blocks of digital systems are defined not analytically, but in the form of subprograms, determine the use of iterative methods for solving equations. The aim of this work is to study the system of multivalued (finite-valued) logical equations from the point of view of the existence of roots and their determination by iterative methods as applied to the methods used for functional-logical modeling of digital systems at the design stage.

II. Conditions for the Existence of Solutions and Their Number

Consider the system:

$$x_i = f_i(x_{1,\dots,}x_n, x_{n+1,\dots,}x_{n+l}), \quad (1)$$

where $x_i, i = 1, \dots, n+l$ are logical variables with the set of variable values $|\mathbf{Z}_i|$;
$x_i, i = n+1, \dots, n+l$ - input variables;
f_i, i = 1,…,n - logical functions with the set of values $|\mathbf{Z}_i|$;
$\mathbf{Z}_i$- a finite set of values of the i-th variable and function, moreover, for each $x_i, i = 1, \dots, n$ in system (1) there is only one equation.

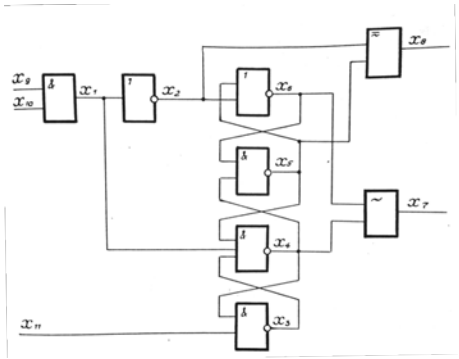
Along with system (1), we will consider its representation in the form of a directed graph $G(\mathbf{V}, \mathbf{E})$, $\mathbf{V} = \mathbf{V}' \cup \mathbf{V}_{in}$, where each vertex $v, v \in \mathbf{V}$ is isomorphic to the variable x_i, i=1,…,n (the set $\mathbf{V}'$) or the input variable x_i, i=n+1,…,n+l (the set $\mathbf{V}_{in}$). The oriented edge e_{ij} is directed from v_i to v_j, if x_i is the argument of f_i. Fig. 1 illustrates the connection diagram of the logic elements of binary logic and the corresponding graph.

For fixed values of the input variables $x_{n+1,\dots,}x_{n+l}$ system (1) has the form:

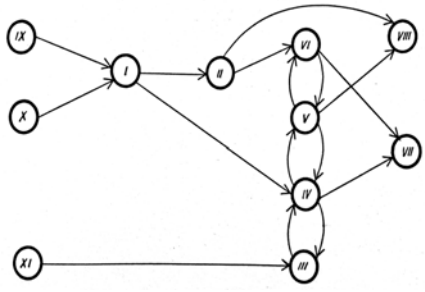
$$x_i = h_i(x_{1,\dots,}x_n), i = 1, \dots, n. \quad (2)$$

If we consider a special case of two-valued logic, then a solution $(x'_1, \dots, x'_n)$ of (2) exists if, for all functions in (2), the values $h_i(x'_1, \dots, x'_n)$ are defined, that is, they are either zero or one. We formulate the conditions for the existence of a solution for (2) in the more general case of the finite-valuedness of variables.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE



a)



b)

Fig. 1. Logical element network (a) and its representation as an oriented graph (b)

For a finite-valued logical function $h_i(x_{1,\dots,}x_n)$ with a set of values $\mathbf{Z}_i = \{z_1, \dots, z_{k_i}\}$ we introduce the Boolean function:

$$eq(h;\, x_{1,\dots,}x_n, z) = \begin{cases} 1 \text{ if } h(x_{1,\dots,}x_n) = z, \\ 0 \text{ if } h(x_{1,\dots,}x_n) \neq z. \end{cases}$$

Theorem 1. For the existence of a solution $x_1 = x_1', \dots, x_n = x_n'$ of the system of finite-valued logical equations (2) it is necessary and sufficient that

$$\bigcap_{i=1}^{n}(\bigcup_{j=1}^{k_i} eq(h_i;\, x_1', \dots, x_n', z_j^i))=1, \qquad (3)$$

where $\{z_1^i, \dots, z_j^i, \dots, z_{k_i}^i\}$ is the set $\mathbf{Z}_i$ of the values of the function h_i.

The number of solutions to system (2) is equal to the number of different sets $(x_1', \dots, x_n')$, for which condition (3) is satisfied.

Evidence. For the truth of the i-th term of conjunction (3) for $x_1', \dots, x_n'$ it is necessary that $eq(h_i;\, x_1', \dots, x_n', x_i')$ =1. Moreover, x_i' can take one of the values of the set $\{z_1^i, \dots, z_j^i, \dots, z_{k_i}^i\}$. If x_i' is equal to one of the values from the set $\{z_1^i, \dots, z_j^i, \dots, z_{k_i}^i\}$, then the disjunction from (3) is true.

For the truth of the conjunction, it is necessary that all members of the conjunction are true, that is, it is necessary that

$$\bigcap_{i=1}^{n}(\bigcup_{j=1}^{k_i} eq(h_i;\, x_1', \dots, x_n', z_j^i))=1. \qquad (4)$$

Therefore, condition (3) is necessary for the existence of solution (2).

Suppose that condition (3) holds on some set $x_1', \dots, x_n'$. Since the conjunction of logical expressions is true if each of them is true, then condition (4) holds for each equation in (2). Therefore, there exists z_j^i such that x_i'=z_j^i, that is, one of the terms on the left-hand side of equality (4) is $eq(h_i; x_1', \dots, x_n', x_i')$, and the latter expression is equal to one. Then, for $x_1 = x_1', \dots, x_n = x_n'$ the i-th equation (2) turns into an identity. The sufficiency of condition (3) is proved.

Due to the sufficiency of condition (3), the number of solutions to system (2) is equal to the number of different sets $x_1', \dots, x_n'$ for which condition (3) is satisfied.

The theorem is proved.

Corollary to theorem. If the functions $h_1, h_2, \dots, h_n$ in (2) can be ordered so that each subsequent function depends only on the values of the previous ones, that is, in the form:

$$h_{i_1}, h_{i_2}(x_{i_1}), h_{i_3}(x_{i_1}, x_{i_2}), \dots, h_{i_n}(x_{i_1}, x_{i_2}, \dots, x_{i_{n-1}}) \qquad (5)$$

then (2) has a unique solution.

Indeed, by virtue of Theorem 1, for this it is necessary to have a unique set $x_{i_1}', \dots, x_{i_n}'$, such that

$$(\bigcup_{j=1}^{k_1} eq(h_{i_1}; z_j^1)) \cdot \dots \cdot (\bigcup_{j=1}^{k_l} eq(h_{i_l}; x_1', \dots, x_{i_{l-1}}', z_j^l)) \cdot \dots \cdot (\bigcup_{j=1}^{k_n} eq(h_{i_n}; x_1', \dots, x_{n-1}', z_j^n)) = 1.$$

Due to the fact that the conjunction of an expression is equal to one if all expressions are equal to one, the validity of the corollary is equivalent to the fact that there is a single set $x_{i_1}', \dots, x_{i_n}'$ such that

$$\bigcup_{j=1}^{k_1} eq(h_{i_1}; z_j^1) = 1; \dots; \bigcup_{j=1}^{k_l} eq(h_{i_l}; x_{i_1}', \dots, x_{i_{l-1}}', z_j^l) = 1; \dots; \quad \bigcup_{j=1}^{k_n} eq(h_{i_n}; x_{i_1}', \dots, x_{i_{n-1}}', z_j^n) = 1. \qquad (6)$$

For fixed values of the input variables, the function h_{i_1} is a constant. Therefore, there is also a single value z_j^1, for which $h_{i_1} = z_j^1$, that is, $eq(h_{i_1}; z_j^1) = 1$.

Let's take x_{i_1}'=z_j^1.

Consider the *l*-th equality from (6). Let the only set $x_{i_1}', \dots, x_{i_{l-1}}'$ be defined from the previous equalities (from the first to the (*l*-1) -th). The function h_{i_l} on this set has some definite value, that is, a single value z_j^l is defined such that $eq(h_{i_l}; x_{i_1}', \dots, x_{i_{l-1}}', z_j^l) = 1$. Let's take $x_l' = z_j^l$. Continuing similar reasoning for the remaining $x_{i_{l+1}}', \dots, x_{i_n}'$, we obtain a single set $x_{i_1}', \dots, x_{i_n}'$.

The corollary is proved.

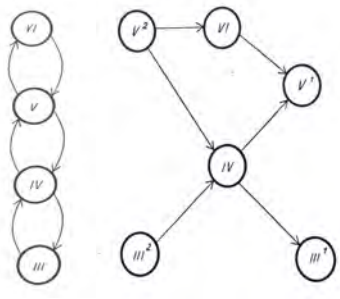
The system of logical equations, which can be ordered in the form (5), corresponds to a structured directed graph without cycles. In this case, the elements of the logical network are rankable, and by successive substitution of expressions for f_j instead of all x_j in $f_i(x_1, \dots, x_j, \dots, x_{i-1})$, all logical variables (1) can be expressed through input variables:

$$x_i = p_i(x_{n+1}, \dots, x_{n+l}), i = 1, \dots, n,$$

where p_i are everywhere defined multivalued logical functions.

Consider the question of the number of solutions to system (2) depending on the structure of the graph $G(\mathbf{V}, \mathbf{E})$. The graph $G'(\mathbf{V}', \mathbf{E}')$ is obtained from the graph $G(\mathbf{V}, \mathbf{E})$ by removing the vertices of the set $\mathbf{V}_{in}$ and the edges outgoing

from them (vertices IX, X, XI, edges IX, I; X, I; XI, III, Fig.1b). System (2) can have a number of solutions different from one only if the graph $G'(\mathbf{V}', \mathbf{E}')$ has cycles. The acyclic part of the graph $G'(\mathbf{V}', \mathbf{E}')$, representing the connection of the cyclic part with the input variables (vertices I, II, edges I, II; II, VI; II, VIII; I, IV, Fig.1b), corresponds to variables whose values are uniquely determined by the state of the input variables. The output acyclic part of the graph $G'(\mathbf{V}', \mathbf{E}')$ (vertices VII, VIII, edges V, VIII; IV, VII, Fig.1b) represents variables whose states are uniquely determined by the states of the network inputs (possibly through the states of the variables the input acyclic part) and the states of the variables of the cyclic part. The states of the variables of the output acyclic part are always defined and do not affect the solution of system (2). We remove from the graph $G'(\mathbf{V}', \mathbf{E}')$ the input and output acyclic parts and obtain the graph $G''(\mathbf{V}'', \mathbf{E}'')$ (Fig.2a).



a) b)

Fig. 2. Cycle part $G''(\mathbf{V}'', \mathbf{E}'')$ of graph $G'(\mathbf{V}', \mathbf{E}')$ (a) and vertex set $\mathbf{V}_c$ = {III, V} dividing onto input and output vertex ones (b)

Let us find in $G''(\mathbf{V}'', \mathbf{E}'')$ such a set of vertices $\mathbf{V}_c$, that after removing the edges outgoing from these vertices, the graph does not contain cycles, and $\prod_{v \in \mathbf{V}_c} |\mathbf{Z}_v|$ is minimal. We transform each vertex $v, v \in \mathbf{V}_c$ into two unconnected vertices v^1 and v^2 in such a way that all the edges entering into v enter the vertex v^1, and all the edges leaving the vertex v are output of the vertex v^2. Then, for given values of the input variables, we can express the variables $x_{c1}^1, \dots, x_{ck}^1$ corresponding to the vertices v^1 of the set $\mathbf{V}_c$, in the form $x_{ci}^1 = q_i(x_{c1}^2, \dots, x_{ck}^2),\ i = 1, \dots, k$. Since the variables x_{cj}^1 and x_{cj}^2 are identical, the solution to system (2) can be replaced by a solution to the system

$$x_i = q_i(x_1, \dots, x_k);\ i = 1, \dots, k;\ \{v_1, \dots, v_k\} = \mathbf{V}_c. \qquad (7)$$

The number of solutions to system (7), and hence to system (2), cannot exceed $\prod_{v \in \mathbf{V}_c} |\mathbf{Z}_v|$. Thus, it is proved

Theorem 2. A solution to system (2) is equivalent to a solution to system (7), and the number of solutions to system (2) does not exceed $\prod_{v \in \mathbf{V}_c} |\mathbf{Z}_v|$.

Let in the example (Fig. 1) the system of equations has the form:

$$\begin{cases} x_1 = x_9 \cdot x_{10} \\ x_2 = x_1 \\ x_3 = \overline{x_4 \cdot x_{11}} \\ x_4 = \overline{x_1 \cdot x_3 \cdot x_5} \\ x_5 = \overline{x_4 \cdot x_6} \\ x_6 = x_2 + x_5 \\ x_7 = x_4 \cdot x_6 + \overline{x_4} \cdot \overline{x_6} \\ x_8 = \overline{x_2} \cdot x_5 + x_2 \cdot \overline{x_5} \end{cases},$$

and $\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{Z}_3 = \mathbf{Z}_4 = \mathbf{Z}_5 = \mathbf{Z}_6 = \mathbf{Z}_7 = \mathbf{Z}_8 = \mathbf{Z}_9 = \mathbf{Z}_{10} = \mathbf{Z}_{11} = \{0,1\}$.

For x_9=1, x_{10}=1, x_{11}=1 x_9 = 1, x_10 = 1, x_11 = 1, the system will have the form:

$$\begin{cases} x_1 = 1 \\ x_2 = \overline{x_1} \\ x_3 = \overline{x_4} \\ x_4 = \overline{x_1 \cdot x_3 \cdot x_5} \\ x_5 = \overline{x_4 \cdot x_6} \\ x_6 = x_2 + x_5 \\ x_7 = x_4 \cdot x_6 + \overline{x_4} \cdot \overline{x_6} \\ x_8 = \overline{x_2} \cdot x_5 + x_2 \cdot x_5 \end{cases} \qquad (8)$$

After determining x_1=1, x_2=0 and taking into account their values in the equations for x_4, x_6, x_8 (removing the input acyclic part of the graph $G'(\mathbf{V}', \mathbf{E}')$) we get:

$$\begin{cases} x_3 = \overline{x_4} \\ x_4 = \overline{x_3 \cdot x_5} \\ x_5 = \overline{x_4 \cdot x_6} \\ x_6 = x_5 \\ x_7 = x_4 \cdot x_6 + \overline{x_4} \cdot \overline{x_6} \\ x_8 = x_5 \end{cases}.$$

The values of the variables x_7, x_8 are determined by the values of x_4, x_5, x_6, that is, the values of the variables of the cyclic part, and do not affect the solution of the resulting system of equations. In this regard, we will consider the solution of the system:

$$\begin{cases} x_3 = \overline{x_4} \\ x_4 = \overline{x_3 \cdot x_5} \\ x_5 = \overline{x_4 \cdot x_6} \\ x_6 = x_5 \end{cases},$$

which corresponds to the graph $G''(\mathbf{V}'', \mathbf{E}'')$ (Fig.2a). Taking $\mathbf{V}_c$ = {III, V}, we get

$$\begin{cases} x_3 = x_3 \cdot x_5 \\ x_5 = x_5 \end{cases}.$$

This system of equations has three solutions:

$$\begin{cases} x_3 = 0 \\ x_5 = 0 \end{cases}; \begin{cases} x_3 = 0 \\ x_5 = 1 \end{cases}; \begin{cases} x_3 = 1 \\ x_5 = 1 \end{cases}.$$

Accordingly, system (8) also has three solutions:

$$\begin{cases} x_1 = 1 \\ x_2 = 0 \\ x_3 = 0 \\ x_4 = 1 \\ x_5 = 0 \\ x_6 = 1 \\ x_7 = 1 \\ x_8 = 0 \end{cases}; \begin{cases} x_1 = 1 \\ x_2 = 0 \\ x_3 = 0 \\ x_4 = 1 \\ x_5 = 1 \\ x_6 = 0 \\ x_7 = 0 \\ x_8 = 1 \end{cases}; \begin{cases} x_1 = 1 \\ x_2 = 0 \\ x_3 = 1 \\ x_4 = 0 \\ x_5 = 1 \\ x_6 = 0 \\ x_7 = 1 \\ x_8 = 1 \end{cases},$$

which can be verified by substituting the given values in (8).

The solution of systems of logical equations can be carried out in various ways. However, the specific form of equations, as well as the fact that when simulating digital systems, block models are set not analytically, but in the form of software modules calculating the values of the function $f_1, \dots, f_n$ from the given values of the arguments $x_1, \dots, x_n$, cause the use of iterative methods of solving [3].

III. SOLVING A SYSTEM OF EQUATIONS BY SIMPLE ITERATION

When solving system (2) by the simple iteration method, the following formulas are used:

$$\begin{cases} x_1^{(j)} = h_1(x_1^{(j-1)}, \dots, x_n^{(j-1)} \\ \vdots \\ x_n^{(j)} = h_n(x_1^{(j-1)}, \dots, x_n^{(j-1)}) \end{cases}; \text{j=1,2,…;} \qquad (9)$$

$x_1^{(0)}, \dots, x_n^{(0)}$ - initial approximation.

System (9) is a unary operation π [14] defined on the set of states $\mathbf{W} = \{(z_{j_1}^1, \dots, z_{j_n}^n) | z_{j_1}^1 \in \mathbf{Z}_1, \dots, z_{j_n}^n \in \mathbf{Z}_n\}$. The graph $H(\mathbf{W}, \mathbf{Q})$ can serve as an iteration model. Operation π defines a set of oriented edges $\mathbf{Q}$ in such a way that the graph has an edge $q(w_k, w_l)$, if $\pi(w_k) = w_l$, where w_k and w_l are vertex labels. The graph under consideration has exactly one edge coming out of each vertex. If *r* is a solution to system (2), then π (*r*) = *r*, and the vertex labeled *r* has a loop. Conversely, if a vertex labeled *r* has a loop, then *r* is the solution to system (2).

The set of collections of functions $h_1, \dots, h_n$ in system (2) is isomorphic to the set of graphs $H(\mathbf{W}, \mathbf{Q})$ with $\prod_{i=1}^n |\mathbf{Z}_i|$ vertices, each of which is incident to exactly one outgoing edge. In this regard, the study of the solution to system (2) can be replaced by the study of the properties of graphs of the indicated form.

A graph $H(\mathbf{W}, \mathbf{Q})$ consists of one or more connected components. In fig. 3 shows an example of such a graph for the case of two-valued logic.

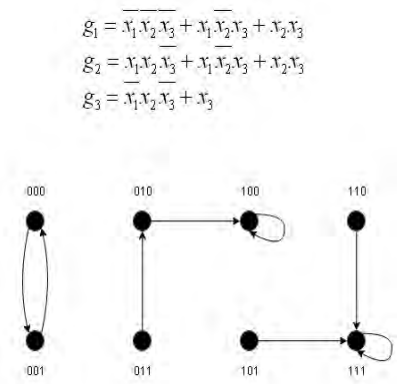


Fig. 3. Graph $H(\mathbf{W}, \mathbf{Q})$ for simple iteration

Each connected component has a cycle reachable from all vertices of the component. If the cycle is a loop, then the component contains the solution; if the cycle is not a loop, then the connected component does not contain a solution. The iteration converges if the initial approximation belongs to a connected component that has a loop. So, when choosing an initial approximation of 101 or 110 (Fig. 3), the iteration converges to the root 111, and when the initial approximation 011 is chosen, to the root 100. In the case of the initial approximations 000 or 001, the iteration does not converge. For the iterations to converge for any initial approximation, it is necessary that each connected component of the graph $H(\mathbf{W}, \mathbf{Q})$ has a vertex with a loop.

Consider the set of operations $\bar{\pi} = \{\pi, \pi^2, \pi^3, \dots\}$, performed on the elements of the set $\mathbf{W}$ for one, two, etc. iterations. The operation π^2 is a transformation $x_i = h_i\big(h_1(x_1, \dots, x_n), \dots, h_n(x_1, \dots, x_n)\big);\ i = 1, \dots, n$. The operation π^3 is a transformation

$$x_i = h_i\big(h_1(h_1(x_1, \dots, x_n), \dots, h_n(x_1, \dots, x_n)\big), \dots,$$
$$h_n\big(h_1(h_1(x_1, \dots, x_n), \dots, h_n(x_1, \dots, x_n)\big);$$
$$i = 1, \dots, n.$$

Thus, the cyclic semigroup $\mathfrak{B} = <\bar{\pi}, \cdot>$ is defined, the generating element of which is π. Since the number of logical functions of *n* finite-valued variables is finite, the pair of indicators $< l_1, l_2 >$ of the semigroup $\mathfrak{B}$ is such that $l_2 > 1$, that is, $\pi^{l_2} = \pi^{l_1}$, where $l_1 < l_2$.

If each connected component $H_j(\mathbf{W}_j, \mathbf{Q}_j)$ of the graph $H(\mathbf{W}, \mathbf{Q})$ has a vertex with a loop, that is, the iteration converges for any initial approximation, then $\pi^l = \pi^{l+1} = \pi^{l+2} = \dots$. That is, starting from π^l, all elements of the semigroup $\mathfrak{B}$ coincide, where *l* is the length of the maximal chain in the graph $H(\mathbf{W}, \mathbf{Q})$, and $\pi^l(w) = r_j$, where *w* is any vertex belonging to $H_j(\mathbf{W}_j, \mathbf{Q}_j)$, and r_j is the root of the component H_j. If the graph $H(\mathbf{W}, \mathbf{Q})$ has one connected component, then $\pi^l(w) = r$ for $w \in \mathbf{W}$, and *r* is the only solution to system (2). Thus proved

Theorem 3. If in the cyclic semigroup $\mathfrak{B}$ defined by (9) $\pi^{l+1} = \pi^l$, then the iteration always converges and the mapping π^l gives a solution depending on the initial approximation. If there is only one solution, then $\pi^l = r$, where *r* is the solution to (2). The value *l* is the length of the maximal chain without loops in the graph $H(\mathbf{W}, \mathbf{Q})$.

Example. Let a mapping π is given by the system of Boolean equations

$$\begin{cases} x_1 = \overline{x_1} \cdot \overline{x_2} + x_1 \cdot x_2 \\ x_2 = \overline{x_2} + \overline{x_1} \cdot x_2 \end{cases}. \qquad (10)$$

Then $\pi^2 = \begin{pmatrix} x_1 \cdot x_2 \\ x_2 + x_1 \cdot x_2 \end{pmatrix}$; $\pi^3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. So, the solution to system (10) is $x_1 = 0,\ x_2 = 1$, the iteration converges for any initial approximation, the length of the maximal chain without loops in the graph $H(\mathbf{W}, \mathbf{Q})$ (Fig. 4) is three.

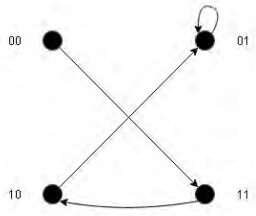


Fig. 4. Graph $H(\mathbf{W}, \mathbf{Q})$ for (10)

When solving the system of equations (2) by the simple iteration method, it is efficient to use the event-driven simulation algorithm. In this case, at each iteration, only those variables are recalculated, among whose arguments there are those that have changed their value.

IV. SOLVING A SYSTEM OF EQUATIONS BY THE GENERALIZED ITERATION METHOD

The iterative solution of system (2) in the general case can be carried out according to the following algorithm:

- take the initial approximation of the values of the variables for the current values;

- for the current values of the variables, calculate the new values of the variables with numbers from $\mathbf{J}, \mathbf{J} \subseteq \{1, \dots, n\}$; update their current value;

- if no solution is obtained, repeat the previous point for a new subset of variables.

The sequence of subsets of numbers of the variables being recalculated $\mathbf{J}^1\mathbf{J}^2 \dots = \{j_1^1, \dots, j_{m_1}^1\}\{j_1^2, \dots, j_{m_2}^2\} \dots$ will be called the trace of iteration. If a part of the trace $\dots \mathbf{J}^k \dots \mathbf{J}^{k+l}$ contains all elements of the set $\{1,\dots, n\}$ and at the last l steps of the iteration no variable has changed its value, then a solution to system (2) is obtained.

Generalized iteration with a trace $\mathbf{JJ} \dots$, where $\mathbf{J} = \{1, \dots, n\}$, is a simple iteration, and with a trace $\mathbf{J}^1\mathbf{J}^2 \dots \mathbf{J}^n\mathbf{J}^1\mathbf{J}^2 \dots \mathbf{J}^n \dots$, where $\mathbf{J}^i = \{i\}$, - Seidel iteration. A multiplicative semigroup is defined on the set $\boldsymbol{\Delta}$ of operations realized under a generalized operation. The elements of the basic set $\boldsymbol{\Delta}$ of the semigroup will be denoted by $\delta^{\mathbf{J}^1\mathbf{J}^2\dots}$, where $\mathbf{J}^1\mathbf{J}^2 \dots = l$ is the trace of the iteration. The multiplication operation is defined by the equality $\delta^{l_1}\delta^{l_2} = \delta^{l_1 l_2}$. A semigroup over $\boldsymbol{\Delta}$ has $2^n - 1$ generators:

$$\delta^{\mathbf{J}} = \begin{cases} x_i, & i \notin \mathbf{J} \\ g_i(x_1, \dots, x_n), & i \in \mathbf{J} \end{cases},$$

where $\mathbf{J}$ is the set of nonempty subsets of $\{1,\dots, n\}$.

On the set of operations $\delta^{\mathbf{J}}$, in turn, an additive semigroup is defined with the operation $\delta^{\mathbf{J}^1} + \delta^{\mathbf{J}^2} = \delta^{\mathbf{J}^3}$, where $\mathbf{J}^3 = \mathbf{J}^1 \cup \mathbf{J}^2$, and generating elements

$$\delta^i = \begin{cases} g_i(x_1, \dots, x_n), & j = i \\ x_j, & j \neq i \end{cases}; i = 1, \dots, n.$$

Each system of equations (2), when it is solved by the generalized iteration method, is isomorphic to a directed graph $Q(\mathbf{W}, \mathbf{K})$ (fig. 5), in which each vertex w labeled $(z_{j_1}^1, \dots, z_{j_i}^i, \dots, z_{j_n}^n)$ incident to $2^n - 1$ outgoing edges, labeled by generating elements $\delta^{\mathbf{J}}$. The edges labeled δ^i can

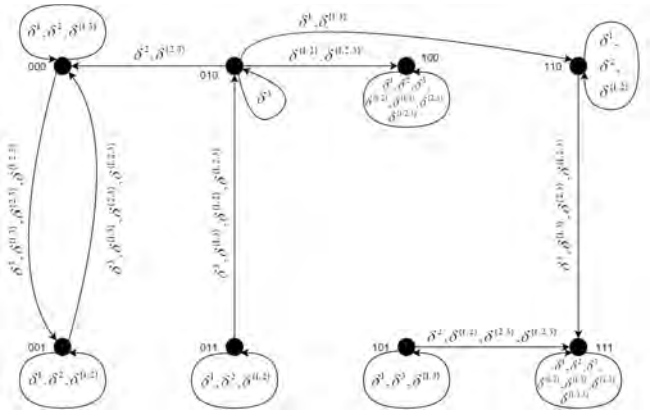


Fig. 5. Graph $Q(\mathrm{W}, \mathrm{K})$ for equations on fig. 3

be either loops or incoming edges of one of $|\mathbf{Z}_i| - 1$ | vertices labeled $\left(z_{j_1}^1, \dots, z_{j_i'}^i, \dots, z_{j_n}^n\right)$, $z_{j_i'}^i \neq z_{j_i}^i$, if $\delta\left(z_{j_1}^1, \dots, z_{j_i}^i, \dots, z_{j_n}^n\right) = \left(z_{j_1}^1, \dots, z_{j_i'}^i, \dots, z_{j_n}^n\right)$. The remaining edges $\delta^{\mathbf{J}}$ are incoming edges of the vertices labeled $(z_{j_1'}^1, \dots, z_{j_n'}^n)$, where $z_{j_i'}^i = z_{j_i}^i$ if $i \notin \mathbf{J}$ and $z_{j_i'}^i = g_i(z_{j_1}^1, \dots, z_{j_n}^n)$ if $i \in \mathbf{J}$. If a vertex $z_{j_1}^1, \dots, z_{j_n}^n$ is a solution to system (2), then all edges outgoing from this vertex with labels δ^i, and hence also with labels $\delta^{\mathbf{J}}$, are loops. And vice versa, if all edges labelled δ^i *and* outgoing from a vertex $\left(z_{j_1}^1, \dots, z_{j_n}^n\right)$ are loops, then $\left(z_{j_1}^1, \dots, z_{j_n}^n\right)$ is the solution to (2). Thus, it is true

Theorem 4. Generalized iteration with a trace $\mathbf{J}^1\mathbf{J}^2 \dots \mathbf{J}^k$ converges at the initial approximation $x_1^{(0)}, \dots, x_n^{(0)}$ if there exists a solution $x_1 = z_{j_1}^1, \dots, x_n = z_{j_n}^n$ to system (2) and a vertex $\left(z_{j_1}^1, \dots, z_{j_n}^n\right)$ in the graph $Q(\mathbf{W}, \mathbf{K})$ is reachable from a vertex with a label $(x_1^{(0)}, \dots, x_n^{(0)})$ along the path $\delta^{\mathbf{J}^1}\delta^{\mathbf{J}^2} \dots \delta^{\mathbf{J}^k}$. A necessary condition for the reachability of a solution from a vertex with a label $(x_1^{(0)}, \dots, x_n^{(0)})$ is that they belong to one connected component.

Of particular interest is the iteration corresponding to synchronous simulation of the ranked circuit. Let a set of feedback variables $\mathbf{X}_c$ be selected, and the structural graph by dividing each vertex $v, v \in \mathbf{V}_c$ into v^1 and v^2 transformed into an acyclic one. Let us rank the vertices $v \in \mathbf{V}$, and, consequently, the equations of system (1), assigning a rank to a vertex v equal to the length of the maximum path from any of the input vertices to the vertex v. Let be $\mathbf{R}_l$ – the set of numbers of the equation (vertices v) with rank l. Consider iterating with a trace $\mathbf{R}_1\mathbf{R}_2 \dots \mathbf{R}_l\mathbf{R}_1\mathbf{R}_2 \dots \mathbf{R}_l \dots$. Obviously, $\bigcup_{l=1}^{\mathbf{L}} \mathbf{R}_l = \{1, \dots, n\}$; $\mathbf{R}_{l_1} \cap \mathbf{R}_{l_2} = \emptyset$ if $l_1 \neq l_2$. Thus, each equation is calculated exactly once during the operation $\delta^{\mathbf{R}} = \delta^{\mathbf{R}_1\mathbf{R}_2 \dots \mathbf{R}_l}$. $\delta^{\mathbf{R}}$ is a generator of a cyclic semigroup on the set of operations $(\delta^{\mathbf{R}})^k, k = 1,2, \dots$. Since the set of values of variables from $\mathbf{X}_c$ is $\prod_{x_i \in \mathbf{X}_c} |\mathbf{Z}_i|$ then $\delta^{\mathbf{R}^l}(\mathbf{X}_c^{(0)}) = \delta^{\mathbf{R}^k}(\mathbf{X}_c^{(0)})$ if $l > k$, and $k \leq \prod_{x_i \in \mathbf{X}_c} |\mathbf{Z}_i| - 1$. Thus, it is true

Theorem 5. If iteration with ranking for a given initial approximation gives solution (2), then each equation must be recalculated no more than $\prod_{x_i \in \mathbf{X}_c} |\mathbf{Z}_i| - 1$ times to obtain a solution.

V. Choosing a Method for Logical Equations Solving in Digital System Simulation

When choosing a method for solving systems of multivalued logical equations in order to obtain quasi-temporal logical diagrams, the following factors must be taken into account.

1. The purpose of the simulation is to test the functions performed by the hardware circuit, to interpret microprograms or software fragments. In this regard, it is important to find the steady-state values of the signals with a minimum expenditure of computer time.
2. The simulation system should use such a method for solving equations, the algorithm of which is universal with respect to the simulated circuits. Such an algorithm should not require human intervention, for example, to select a set of feedback loops.
3. A distinctive feature of digital systems is the bidirectionality of buses and lines, which leads to a change in the structural graph $G(\mathbf{V}, \mathbf{E})$ depending on the internal state of the system, and, as a consequence, to a change in the ranks of components during the simulation process.
4. When designing digital system hardware, developers widely use the fact that the response times of blocks are not zero and have finite values. Thus, the use of a pipelined register in digital systems based on microprogramming makes it possible to combine in time the execution of the current micro-instruction and the selection of the next micro-instruction from the ROM. In this case, the

execution of the micro-command is carried out during the delay interval for issuing the next micro-command from the ROM relative to the moment of supplying its address.

When simulating equipment on blocks of a low degree of integration, the solution of equations by the Seidel iteration method with quasi-disruption of feedback loops and ranking of equations is effectively used . Physically this corresponds to zero-latency synchronous simulation. This minimizes the time spent on calculating the feedback. Indeed, with known feedback signals for calculating the values of the circuit signals, the ranging iteration is most efficient, since each signal in this case is calculated only once. However, the automatic selection of feedback loops and the ranking of components require additional computer time. In the case of using blocks with bidirectional outputs, when the internal states of such blocks change, it is necessary to re-calculate the ranks of the components.

The event-driven algorithm of simple iteration is somewhat more costly, since even with known feedback signals, the value of a number of variables has to be recalculated more than once. However, in this case, there is no need to perform component ranking; the bidirectional nature of block pins also does not lead to additional complications. The event-driven algorithm of simple iteration is universal with respect to the scheme of the simulated hardware.

In addition, simple iteration assumes equal delays of hardware blocks, which corresponds to the intuitive idea of developers when drawing up a schematic diagram.

Based on the above, and also taking into account the experience of creating and operating simulating systems at the functional-logical level, specified by the block connection scheme, which provides quasi-temporal logical diagrams, we can conclude about the effectiveness of using the event-based algorithm of simple iteration in this case.

The algorithm for solving systems of logical equations in the simulating system has the form.

1. Take as an initial approximation the values of logical variables and the state of the internal registers obtained in the previous cycle.
2. Determine the input signals at the given cycle.
3. Refer to the m_i models, whose inputs are connected to the changed signals at the nodes, calculate the new values of the internal variables r_i (the state of the internal registers) and the output signals of the m_i models. If a combination of input signals applied to any model in combination with an internal state is prohibited, then the model is set to an error state; simulation stops.
4. If the number of changes in the input signal of any model exceeds the specified one, then stop the simulation by looping, issuing a message about the absence of a solution to the system (2).
5. For all circuit nodes to which the outputs of models with changed signals are connected, calculate new logical values in the node. If a logical value of the signal is not defined for any node, display an error message in block merging.
6. If for any nodes of the circuit the values of the logical values of the signals have changed, then go to step 3. If all the values have remained unchanged, then the solution (2) is found.

Thus, the filing of prohibited signal combinations on the block model is detected at step 3 of the algorithm; the absence of steady signals at the outputs of the blocks, that is, the absence of a solution to system (2), at step 4 of the algorithm, errors in combining blocks - at step 5 of the algorithm. Errors in the performance of the functions of hardware, as well as in the software and firmware are detected by the developer by analyzing the values of logical signals and states of the internal registers of blocks at each step of the simulation.

VI. Conclusion

A theoretical analysis of iterative methods for solving finite-valued, in a particular case, binary systems of logical equations is carried out. The conditions for the existence and determination of the number of solutions are considered. Different iterative methods correspond to different engineering approaches in functional-logical simulation of digital systems design. The conducted research makes it possible to determine the correspondence between the mathematical results and situations arising in the simulation of designs of digital systems.

References

[1] P. Keresztes, A. Tukacs, and M. Török, “A multi valued logic VHDL package for switch level simulation of novel digital CMOS circuits”, 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), Bhubaneswar, India, 2018, pp. 25-28.

[2] A. Bara, P. Bazargan-Sabet, R. Chevallier, E. Encrenaz, D. Ledu, and P. Renault, “Formal verification of timed VHDL programs”, 2010 Forum on Specification & Design Languages (FDL 2010), Southampton, 2010, pp. 1-6.

[3] A. Ivannikov, A. Romanov, and A. Stempkovsky, “Set-theoretic Model of Digital Systems Functioning”, IEEE International Siberian Conference on Control and Communications (SIBCON), Moscow, 2016. Article number: 308fu4t.

[4] S. Kunapareddy, S.D. Turaga, and S.S.T.M. Sajjan, “Comparision between LPSAT and SMT for RTL verification”, 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], Nagercoil, 2015, pp. 1-5.

[5] A. Ivannikov, N. Levchenko, and I. Romanova, "Formal Description of Possible Input Logical Signal Data Sequences for Digital Systems and Their Blocks," 2018 IEEE East-West Design & Test Symposium (EWDTS), Kazan, 2018, pp. 1-5.

[6] Y. Tai, W. Hu, Guo Lantian, B. Mao, and D. Mu, “Gate level information flow analysis for multi-valued logic system”, 2017 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, 2017, pp. 1102-1106.

[7] S.C. Mane, S. P. Hajare, and P. Dakhole, “Current mode quaternary logic circuit”, 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2017, pp. 0825-0829.

[8] A.P. Sooriamala, and E. Poovannan, “Synthesis of multiple valued logic digital circuits using CMOS gates”, 2017 International Conference on Innovations in Electrical, Electronics, Instrumentation and Media Technology (ICEEIMT), Coimbatore, 2017, pp. 383-388.

[9] N. N. Prokopenko, N. I. Chernov, V. Yugai, and N. V. Butyrlagin, “The element base of the multivalued threshold logic for the automation and control digital devices”, 2017 International Siberian Conference on Control and Communications (SIBCON), Astana, 2017, pp. 1-5.

[10] K. Shimabukuro, and M. Kameyama, “Fine-grain pipelined reconfigurable VLSI architecture based on multiple-valued multiplexer logic”, 2017 IEEE 47th International Symposium on Multiple-Valued Logic (ISMVL), Novi Sad, 2017, pp. 19-24.

DRAM structure with prioritized memory bank using multi-VT bit cells architecture

Narek Mamikonyan
Microelectronics
Yerevan State University
Yerevan, Armenia
narekmamikonyan073@gmail.com

***Abstract*—Increasing demand for bigger dynamic random-access memory (DRAM) memory capacitance, rises the power consumption of memory. Higher power consumption brings to chip heating, memory work uncertainty, IR on the chip, etc. Along with this, memory access time is also one of the important points in the chip working. For power saving propose, low-power cells are used like High threshold voltage (VT) standard cells. However mentioned cells have high transition time, which brings to memory access time increase.**

This paper represents Multi-Vt bit cell memory architecture, which is combined with the memory bank prioritizing algorithm, which is aimed to solve upper mentioned issues. Prioritizing of memory is based on counting each memory address access times and assigning unique cost number to them. Later this cost value is used to reconfigure memory addresses to have appropriate threshold voltage (VT) memory cell where the address is commonly used. The memory address mapper block is used For controlling memory block addressing.

Testing results show, that using proposed method power consumption reduced by 5% compared with the standard Low VT memory bank. However, as control logic is used to switch between memory banks, the overall logical cell count is increased by 10%. Logic cell count increase brings an additional area increase in the design.

Keywords—multi-VT, bit cell, memory bank, low-power, low-voltage

I. INTRODUCTION

Technology scaling and memory capacitance increase demands bring additional restrictions and complexity for the memory design process [1]. Power consumption reduction and memory access time are equally important for memory design. DRAMs are commonly used for their small sizes and low latency. But DRAMs have high power consumption caused by the periodical refresh processes[2]. For particular DRAM, the refresh process can be determined in a way, to have a golden mine between power consumption and timing characteristics[3]. Moreover, standard cell VT can help to solve the power consumption issue but it will cost degradation of the timing characteristics. Memory bank information at a particular address can be accessed at different times in some period. If the data is accessed frequently it is good to have short access time for that address, but on the other hand, a possible scenario is to have another address in the same bank which is rarely accessed. It is better to select memory by determining which accumulative frequency of total addresses access count, but in that case, the memory will have addresses that are not passing timing criteria. To meet timing criteria best timing performer memory cells must be used for frequently accessed addresses. Memories with high frequency will have power overconsumption. To solve the upper mentioned issues, swapping memory cells for whole memory array from one VT to another VT can be used. It will not be the best optimal way to use resources. This paper presents another method to optimize power consumption by using multi-VT DRAMs among additional prioritizing logical blocks. The prioritizing logical block is used for address mapping for one VT memory bank to another.

II. TYPES OF MEMORIES

In Very Large Scale Integrated (VLSI) circuit standard cell libraries are the most commonly used libraries. They are optimized libraries, hence using them can fasten the design process. For different purposes, standard cells can be in different VTs, hence libraries for them are called the exact "VT" library.

The standard cells with High threshold voltage consume low power, but however, they have lower performance. In contradiction to high threshold voltage cells, low threshold voltage cells consume more power, but they perform better in terms of timing. In designs, usually consumed small power is more desirable[4]. If cells of low threshold voltage are swapped to high threshold voltage cells, in design significant reduction of power usage can be achieved while depredating timing parameters by a small amount.

Usually, different threshold voltage cells have the same size and footprint, hence swapping them between each other in design, can be reached without any issues.

As experimental results show, using different threshold voltage cells in the same design and doing proper optimization, one can achieve up to ~2 times leakage power reduction, which will bring to area variation with non-significant value (about 3 %).

Memory cells can consist of different threshold voltage transistors. Some threshold voltage cells are for saving power, while others are for improving timing. In memory design, cells of the memory block need to be selected based on the application and restriction on the chip[5]. Most commonly used types for memory cell threshold voltages are:

- Low-voltage memory cells
- Low-power memory cells

Low voltage memory cell is structured by low voltage (LVT) standard cells. Transistors of these cells are with thin oxide, hence their threshold voltage to the gate is lower [6]. This is giving the advantage of shorter switching time, but sub-threshold currents are higher, which is bringing additional power consumption.

Low power memory blocks are created using high VT standard library cells . Transistors of mentioned cells are with

thick oxide, and the voltage applied to gates is higher. These are giving the advantage of limiting current and decreasing consumed power [7]. However, these cells are slow compared with LVT cells.

Memory address prioritizer is a logic unit that counts memory address accesses per bank by giving them costs. Costs will be later used for selecting correct VT memory cells for minimizing power consumption and keeping timing criteria in their limits.

Each memory bank from memory block has its counter. The addresses from the same bank that have been accessed are increasing counter value. Comparator module recalculates priorities of the memory banks after an exact amount of accesses.

Block diagram of the proposed logic unit consists of 4 main parts "Fig. 1":

- Memory address counter
- Comparator module
- Prioritizing module
- Memory address mapper
- Memory banks

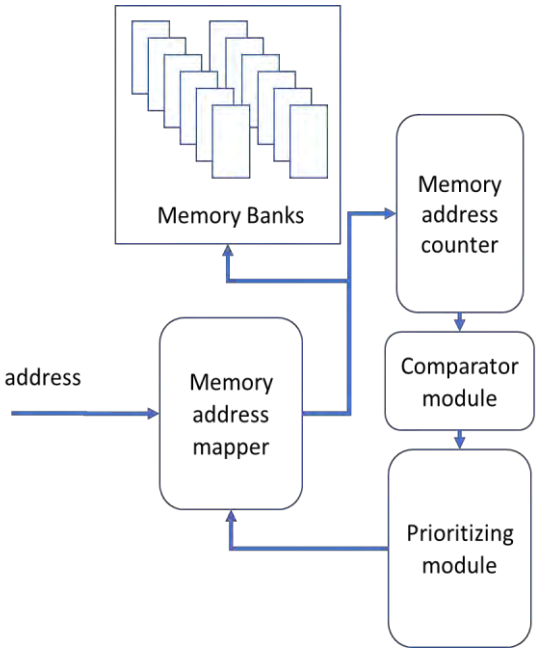


Fig. 1. Memory address prioritizer block diagram

The memory address counter is a block which is responsible for counting memory address usage times. Each memory bank with different VT has a separated counter for all memory addresses. Any time the maximum count of calls is done to the memory bank system, a comparator for different memory VT call count is activated by a signal and starts comparison for getting most called addresses. After this, the sorted cost list for each VT type memory is transferred to the prioritizing module. The prioritizing module makes a decision for VT bank mapping and refreshes information in the swapper module to make proper swapping for rearranging memory addresses. Swapper module, using Advanced memory swapping algorithm, to change memory information from one VT bank to another VT bank.

III. Advanced Memory Swapping Algorithm

Advanced Memory swapping algorithm is for information swapping between different memories (VT memories). The main components of memory swapper are swap module and address serializer "Fig. 2".

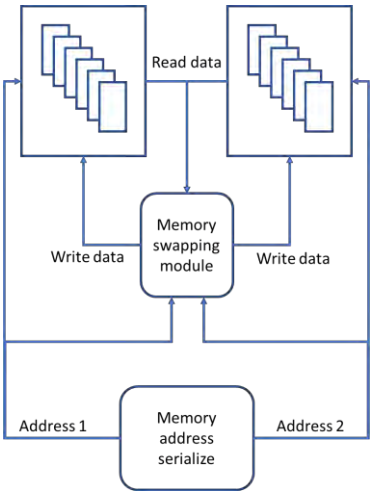


Fig. 2. Memory swapping algorithm block diagram

When an advanced memory swapper is getting proper command from a prioritizer, it starts the process of memory swapping. At first, is it getting addresses of memories from different banks and serially transfers one by one addresses to the memory swapping module and set addresses for the correct bank.

After getting addresses, the memory swapper module sends a read signal to two memory banks. As a result memory swapping module is getting information from two addresses. Once data is obtained, the memory swapper is sending write to enable signal to memory banks with the swapped information. This process continues until memory address serializer is fully past through all addresses of memory banks.

As a result of the overall process, all data from one bank will be transferred to another bank.

IV. Advanced Memory Swapping Algorithm

For demonstrating the proposed method, 2 types of memories (3T 8GB DRAM, 3T 16GB DRAM) with 2 contradictive VTs (Low-voltage, Low-power) are used. Technology for designs is selected to be 32/28 nm flat technology with a Typical process, 0.78V Voltage, 125oC Temperature.

Upper mentioned modules are described with Verilog HDL and synthesized to gate-level representation. Tests are done on two critical cases:

- All addresses are equally used
- Half of all addresses are frequently used, others are rarely used

- Frequent address use is being changed by the time the way, that at each special period of time all addresses of one VT memory are most used. A special period of time is selected to be aligned with the prioritizer update time.

Based on testing results "Table 1", the worst case is being obtained when the upper described third scenario is used.

Results are expected, as in the mentioned case, after each prioritizer update, address frequencies are being changed in the opposite direction of prioritizer decision.

Also, it is noteworthy, that in the case of the first scenario, prioritizer will not be able to take control over the address changing process, as all memory addresses are equally used, thus there is not any advantage in memory swapping.

The best performance can be obtained in scenario two, where addresses are being distributed based on used address frequency.

TABLE I. TESTING RESULTS

Memory Density	Power consumption estimation (scenario 1)	Power consumption estimation (scenario 2)	Power consumption estimation (scenario 3)	Standard Dram logic Standard cell count	Proposed method logic standard cell count
8 Gb	374 mW	327 mW	442 mW	1147	1273
16 Gb	619 mW	542 mW	749 mW	1874	2062

V. CONCLUSION

As memory power usage and timing criteria are tightened, memory control mechanisms are being updated time by time.

In this paper memory, the bank control method is described, which is helping to change memory addresses dynamically by the time of memory evaluation. Memory addresses with higher call count are swapped to faster memory banks of low-voltage cells.

Results show, that with the proposed method and inappropriate memory usage condition, power consumption is being reduced by 5%. However, as control logic is enlarged, the area of overall logic block increases by 10%.

REFERENCES

[1] Prateek Asthana et.al, "Performance Comparison Of 4T, 3T AND 3T1D DRAM Cell Design on 32nm Technology", JSSATE, ICCSEA, SPPR, VLSI, WiMoA, SCAI, CNSA, WeST - 2014, pp. 121–133, 2014.

[2] Soliv, Viplav. (2011). Using CMOS Sub-Micron Technology VLSI Implementation of Low Power, High Speed SRAM Cell and DRAM Cell. International Journal of VLSI Design & Communication Systems. 2. 143-153. 10.5121/vlsic.2011.2412.

[3] Y. Sato *et al*., "Fast cycle RAM (FCRAM); a 20-ns random row access, pipe-lined operating DRAM," *1998 Symposium on VLSI Circuits. Digest of Technical Papers (Cat. No.98CH36215)*, Honolulu, HI, USA, 1998, pp. 22-25, doi: 10.1109/VLSIC.1998.687990.

[4] Itoh K. (2002) Trends in Ultralow-Voltage RAM Technology. In: Hochet B., Acosta A.J., Bellido M.J. (eds) Integrated Circuit Design. Power and Timing Modeling, Optimization and Simulation. PATMOS 2002. Lecture Notes in Computer Science, vol 2451. Springer, Berlin, Heidelberg

[5] Pranita J. Giri, Sunanda K. Kapde, "Implementation of DRAM Cell Using Transmission Gate" in International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Issue 4, April 2017.

[6] M. Srivastav, S. S. S. P. Rao and H. Bhatnagar, "Power reduction technique using multi-Vt libraries," *Fifth International Workshop on System-on-Chip for Real-Time Applications (IWSOC'05)*, Banff, Alberta, Canada, 2005, pp. 363-367, doi: 10.1109/IWSOC.2005.92.

[7] Yen-Te Ho and Ting-Ting Hwang, "Low Power Design Using Dual Threshold Voltage," Design Automation Conference, 2004, pp. 205-208.

IR drop estimation and optimization on DRAM memory using machine learning algorithms

Narek Mamikonyan
Microelectronics
Yerevan State University
Yerevan, Armenia
narekmamikonyan073@gmail.com

Nazeli Melikyan
Microelectronics
National Polytechnic University of Armenia
Yerevan, Armenia
nazeli@synopsys.com

Ruben Musayelyan
Microelectronics
National Polytechnic University of Armenia
Yerevan, Armenia
musayely@synopsys.com

***Abstract*—The IR drop effect on the supply network increases very fast with technology scaling. This can lead to an integrated circuit (IC) speed reduction and timing issues. For IR drop effect reduction, different techniques are used, such as decoupling capacitor insertion, wire sizing, etc.**

In DRAM memories also, the IR drop effect plays a significant role, as aside from timing issues, in DRAM, IR drop can lead to more issues, up to stored-value loose.

In this paper, an approach to IR drop estimation on DRAM memory is used, which is providing data about the possible issue on memory banks using a machine learning algorithm and provides solutions on IR drop reduction. Testing results show, that with using the proposed method, IR drop reduces by 13%, but the available routing track count is decreased by 14%.

Keywords—IR drop, DRAM memory, regression algorithm

I. INTRODUCTION

Daily increasing technology rises the demand for having reliable power and ground (PG) networks. Scaling technology brings a decrease in metal width, which is directly connected with metal resistance. High resistance will affect on the IR drop, increasing the effect of it on the supply network [1].

Possible fixing solutions of IR drop violations are adding more power and ground nets on design. These will ensure a bigger intersection are between two metal layers, which decreases the current on the exact node and, as a result, decreases IR drop [2].

On the other hand, dense power and ground network blocks available routing tracks, which is affecting to routing resources [3]. Hence in dense designs, the trade-off between dense power network and free enough routing tracks for routing must be kept [4]. For achieving this result, some techniques are used, that are mainly applicable for standard cell designs.

For DRAM memories, IR drop can cause issues, which can lead to inaccurate work of memory [5]. DRAM memories are being refreshed time by time (refresh cycle) and having less voltage caused by IR drop, can bring to miss stored 1bit data on the memory cell.

For resolving upper described issues on memories, different techniques are used [6]. Among these techniques, some [7] are using sizing power and ground network, which can affect to routing resources. Others [8] are using the special placement methodology, for ensuring a limited IR drop on power and ground rails. Aside from the mentioned solutions, partitional IR drop fix is decided to be the best solution so far. In this technique, initial power and ground network is being created, while keeping available routing tracks and after that, more power and ground nets are being added on IR drop violated places.

In this paper, another approach to IR drop fixing is shown. The main purpose of the demonstrated method is to have more routing tracks while keeping the IR drop in limits. As memory pins are dense, more routing tracks are needed for properly connecting memories, otherwise, a big count of detours will bring timing violations on memories. After violated area extraction, local power and ground connections are added using via-ladder insertion on the memory rails. For proper estimation of violated areas, machine learning methods are used.

After a successful compilation of the proposed method, structure “Fig. 1” with via ladders and power and ground network is being tested again on IR drop. If IR drops in previous violated areas are respective to desired values, the process is being finished. Otherwise, if there are still violated areas, some more local changes are being done iteratively. Iterations of the process can be controlled by hard values, or by IR drop desired limits.

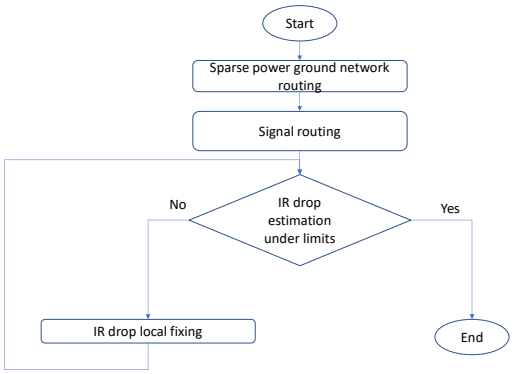


Fig. 1. IR drop fixing method

II. VIA-LADDER DEFINITION AND TECHNIQUE

For advanced node low-threshold transistors, variation is becoming non-Gaussian, hence it makes estimation of timing and power difficult while placing and routing.

The main issue with timing and power is being seen because of metal and interconnect (via) resistance [9], which is increasing with technology scaling and narrow metal/via usage. As a solution to this, special structures of metal/via connection is used, also known as via ladder or via pillar. This structure is a layer promotion method, in which frequent signal routes are shifted to upper metal layers, to use the advantage of high metal layers’ low resistance. With using

via pillar, the current is distributed among more metal shapes/layers, hence voltage drop on one shape is decreasing because of smaller current.

The via pillar "Fig. 2" is a conventional structure to standard redundant vias. It is using closely spaced via pairs for each metal layer from the bottom of the pillar to the up. All layers are in their preferred directions. The structure is forming a lattice configuration, which is passing through several layers.

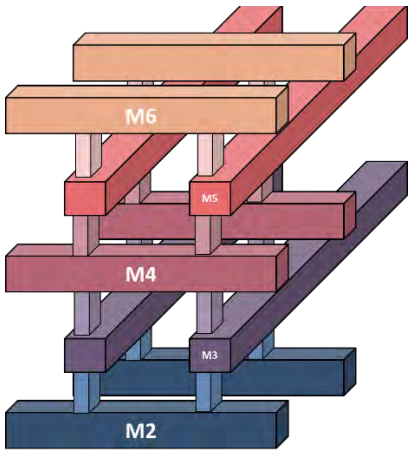


Fig. 2. VIA pillar

Many places and routing tool have emended via ladder insertion modules, which is performing via ladder insertion after the placement stage. As experiments show, with using via ladders, the performance of the design is being improved by ~9% in comparison with standard redundant vias.

III. IR DROP ESTIMATION

In the IC design supply network is distributed uniformly through metal/via layers. As metal layer resistance is finite, the passing current creates a voltage drop on it. This phenomenon is called an IR drop. IR drop on rails can lead to bouncing on power and ground network, particularly can decrease the voltage of power rail and increase the voltage on the ground rail. "Fig. 3"

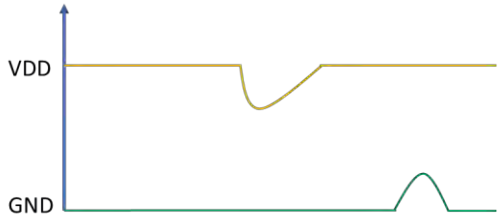


Fig. 3. Power network bounce

Based on Ohms law (Equation 1), the unexpected voltage drop can accrue when unexpected current is passing through high resistance wire. Because of this, the required voltage is not reaching to the memory cell. This is leading to issues of low performance and stability.

$$V = IR \quad (1)$$

Via ladder insertion methodology described in section 1, can help to prevent some amount of IR drop by increasing metal-via connections, hence decreasing resistance of connection node.

IV. IR DROP ESTIMATION AND OPTIMIZATION ON DRAM MEMORY

In this paper proposed algorithm is aimed to solve voltage drop across DRAM memories. As DRAM memory needs a high amount of signal routes, hence for keeping available routing tracks, regular sparse PG network is being created. After the successful signal routing of memory connections, the pre-trained machine learning regression model is estimating the possible IR drop on the PG supply network. As a result of model estimation, areas with predicted IR drop are generated. In the appropriate areas, based on drop value, via ladders are added.

For model training, metal and via sheet resistance and memory draw current is used as input information.

The main disadvantage of the regression model is for each metal stack and technology node the model must be retrained properly, to meet metal layer resistance for that stack/technology.

Along with this, because of the high count of metal layers and a small amount of model training set, the accuracy of perdition can variate.

V. VIA LADDER INSERTION ON DRAM RAILS

Using section 2 described IR drop estimation results, the proposed method is generating areas, in which most IR drop is accrued and generates proper via-ladders to overcome high IR drop. Generated Via ladders are intended to be dense enough to ensure the required IR drop while keeping already existing signal routing.

Overall, generated via-ladders can be in 3 types "Fig. 4" and each of them has its advantages and disadvantages.

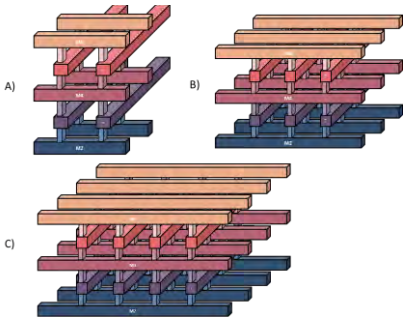


Fig. 4. 3 different types of VIA ladder

VI. TESTING RESULTS

The proposed method is tested on 2 types of 3T DRAM memories. The process for transistors is selected to be typical-typical, the voltage on rails is 0.76V, working temperature is 25ºC.

Results (Table 1) show, that with using the proposed method, the total IR drop has been reduced by 13%. Metal total length increased with 14% which is acceptable, as it does not affect to pre-routed signal nets. Overall elapsed time with using the proposed method, without considering the regression model training process, increases by 11%.

TABLE 1 TESTING RESULTS

Type	CPU runtime	Max IR drop	Total metal length mkr
Sparse power/ground	**3812 s**	**1.5%**	**183.4**
Proposed method	**4265 s**	**1.3%**	**209.1**

CONCLUSION

In this paper, another approach to power and ground network IR drop estimation and reduction is proposed. The method is using a machine learning regression model to predict possible IR drop values and adds via ladders on violated areas.

Results show that, with using the proposed method, total IR drop value can be decreased by 5%, however, total elapsed time for the run is increased. Along with time increase, metal shape count is also being increased, but this is not leading to any unexpected issues, as via ladder insertion is being done with signal route aware method. One more point is that for proper work of demonstrated method, the machine learning model is used, this adds need to be pre-trained the model for each technology/metal stack.

REFERENCES

[1] Y. Fang, H. Lin, M. Sui, C. Li and E. J. Fang, "Machine-learning-based Dynamic IR Drop Prediction for ECO," 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), San Diego, CA, 2018, pp. 1-7, doi: 10.1145/3240765.3240823.

[2] S. Lin *et al*., "IR drop prediction of ECO-revised circuits using machine learning," *2018 IEEE 36th VLSI Test Symposium (VTS)*, San Francisco, CA, 2018, pp. 1-6, doi: 10.1109/VTS.2018.8368657.

[3] X. Jian, P. K. Hanumolu and R. Kumar, "Understanding and Optimizing Power Consumption in Memory Networks," 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), Austin, TX, 2017, pp. 229-240, doi: 10.1109/HPCA.2017.60.

[4] Pant, S., Blaauw, D. and Chiprout, E. (2007) Power Grid Physics and Implications for Cad. IEEE Design and Test of Computers, 24, 246-254.

[5] Schmidt, Daniel , Wehn, Norbert. (2009). DRAM power management and energy consumption. Proceedings of the 22st Annual Symposium on Integrated Circuits and Systems 10.1145/1601896.1601937.

[6] T. Wu, S. Gharahi and J. A. Abraham, "An area efficient on-chip static IR drop detector/evaluator," 2009 IEEE International Symposium on Circuits and Systems, Taipei, 2009, pp. 2009-2012, doi: 10.1109/ISCAS.2009.5118186.

[7] Khalil, D.E. and Ismail, Y. (2006) Optimum Sizing of Power Grids for IR Drop. Proceedings of the IEEE International Symposium on Circuits and Systems, Island of Kos, Greece, 21-24 May 2006, 481-484.

[8] Kahng, A.B., Liu, B. and Wang, Q. (2007) Stochastic Power/Ground Supply Voltage Prediction and Optimization via Analytical Placement. IEEE Transactions on Very Large Scale Integration (VLSI) Circuits, 15.

[9] Lu, Lee-Chung. (2017). Physical Design Challenges and Innovations to Meet Power, Speed, and Area Scaling Trend. 63-63. 10.1145/3036669.3038255.

Bit Depth Impact Analysis of the Gaussian Process Quantization Errors

Aleksey S. Gvozdarev, *Member*, *IEEE*,
Infocommunication and radiophysics department,
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
a.gvozdarev@uniyar.ac.ru, ORCID: 0000-0001-9308-4386

Yury A. Bryukhanov
Infocommunication and radiophysics department,
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
bruhanov@uniyar.ac.ru

***Abstract*— The research considers the problem of the statistical description of the analog-to-digital converter's output signal power distortion due to nonlinearity of the inner quantization device characteristics. The quantizer is assumed to be driven by the input zero mean circular symmetric Gaussian process. The quantitative and qualitative analysis of the power suppression depending on the bit depth of the ADC is performed. The output signal variance is derived for two number representation forms: direct and complementary (both with truncation), and two types of ADCs overloading strategies: saturation and reset. The obtained results are compared to the idealized quantization device with infinite characteristic (no overloading case) and classic linear quantization model with a uniform error distribution.**

***Keywords*— nonlinearity, quantization, power distortion, ADC, Gaussian process, bit depth**

I. INTRODUCTION

Nowadays modern trends for ad-hoc communication system design, for instance, software-defined radio (SDR) [1-2], include a concept of shifting digital devices of the receiving circuitry (e.g., analog-digital converters (ADC)) as close as possible to the antennas' output with the consequent replacement of analog elements and assemblies with digital signal processing units. The practical adoption of such a methodology leads to an essential demand of high-resolution quantization devices (being the core element of the ADCs) since the quality of reproduction of an input analog signal by its digitized version is generally associated with the ADCs' bit depth.

However, for various modern communication standards (including 5G [3-4]) their number greatly increases, limiting the possible applications. This is mainly due to the fact that most of those standards adopt the Massive MIMO technology [5] as one of the corner-stone elements. Its deployment leads to the dramatic increase in ADCs' number since they should be used independently for every receiving channel. This causes an almost exponential increase in power consumption and cost [6].

The feasible solutions of that problem include the possible reduction of the ADCs' resolution [7] (up to one bit), increasing their number to keep the necessary transmission rate or use the combination of high and low resolution ADCs [7].

In both cases, the reduction of the bit depth sufficiently changes statistical properties of the output signal (because of the greater impact of the quantizer characteristics' non-linearity). In this case, the classic linear statistical model [8-9] is no longer applicable. And since the probabilistic description of the output signal greatly influences the adopted detection or demodulation strategy [10] its analysis is of primary importance for the receiver design.

The objective of the present research is the statistical analysis of the quantization effects depending on the assumed number representation form and bit depth. The quantization devices' impact is described in terms of the output signal power distortion (referenced to the power of the Gaussian input process).

II. SYSTEM MODEL

Let us assume that from the statistical viewpoint the input signal $\xi(t)$ of the ADC's *R*-bit quantizer, sampled with a proper sampling time interval Δt, can be represented as a zero-mean circular symmetric Gaussian process with variance σ_x^2, i.e. $\xi_i \sim \mathcal{N}(0, \sigma_\Sigma)$, where $i = \lfloor t / \Delta t \rfloor$ – is the number of the discrete sample. Since the research mainly concentrates upon the power distortion of the output signal, henceforth only the first-order statistical description is adopted. This broadens the applicability of the derived results since they are valid for arbitrary input signal possible correlation models.

As for the relevance of the assumed input signal model, it can be interpreted implicitly as the desired signal, as a pure noise or as a mixture of signal and noise.

It should be also noted that for a wide range of ad-hoc modulations (for example high-bit QAM, varieties of OFDM, etc. in case of communication systems of the 5th generation) samples of signals with the complex signal-code design under specific conditions can be successfully assumed as Gaussian random variates within the framework of the probabilistic-statistical approach [11-13].

Throughout the paper, we'll assume that the inside the ADC the quantizers' instantaneous output level is being represented in either direct of complementary code [9], with truncation for both cases.

Moreover, it should be noted that the real-life ADCs are sensitive to overload. To cope with this problem two strategies are adopted: overload with the saturation (typical for classic quantizers [14]) or with the reset (typical for Δ-Σ quantizers [15]). The two situations are supplemented by the idealized quantizer with no overload.

Hence combining all of the above models, the following input-output quantizers' characteristics $\eta = Q(\xi)$ are analyzed:

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

A. Direct code representation with overload saturation (model № 1):

$$\eta_{dir_1} = Q_{dir}\left(\xi\right) = \begin{cases} q\left\lfloor \dfrac{\xi}{q} \right\rfloor, & \left|\xi\right| \le 1 \\ \operatorname{sgn}\left(\xi\right), & \left|\xi\right| > 1 \end{cases} \quad (1)$$

where q stands for the quantization step and is related to the ADCs' bit depth R as $q = 2^{-R}$, $\lfloor \bullet \rfloor$ is the truncation operator and $\operatorname{sgn}(\bullet)$ is the classical sign function.

B. Direct code representation with overload reset (model № 2):

$$\eta_{dir_2} = Q_{dir}\left(\xi\right) = \begin{cases} q\left\lfloor \dfrac{\xi}{q} \right\rfloor, & \left|\xi\right| \le 1 \\ 0, & \left|\xi\right| > 1 \end{cases} \quad (2)$$

C. Direct code representation with idealized quantizer (model № 3):

$$\eta_{dir_3} = Q_{dir}\left(\xi\right) = q\left\lfloor \frac{\xi}{q} \right\rfloor \quad (3)$$

D. Complementary code representation with overload saturation (model № 1):

$$\eta_{comp_1} = Q_{comp}\left(\xi\right) = \begin{cases} q\left\lfloor \dfrac{\xi+1}{q} \right\rfloor - 1, & -1+q \le \xi < 1-q \\ 1-q, & 1-q \le \xi \\ -1, & \xi < -1+q \end{cases} \quad (4)$$

E. Complementary code representation with overload reset (model № 2):

$$\eta_{comp_2} = Q_{comp}\left(\xi\right) = \begin{cases} q\left\lfloor \dfrac{\xi+1}{q} \right\rfloor - 1, & -1+q \le \xi < 1-q \\ 1-q, & \left(1-q \le \xi\right) \vee \left(\xi < -1+q\right) \end{cases} \quad (5)$$

F. Complementary code representation with idealized quantizer (model № 3):

$$\eta_{comp_3} = Q_{comp}\left(\xi\right) = q\left\lfloor \frac{\xi+1}{q} \right\rfloor - 1 \quad (6)$$

It should be noted that in practice the adopted model (*A* to *F*) is chosen based on some existing prior information about the input signal. For instance, in case when $\xi(t)$ represents a jammer (or a clutter) the reset option is favourable. On the other hand, when $\xi(t)$ is a desired signal it reasonable to give preference to saturation. Models *C* and *F* are mainly used as a benchmark.

III. Derived results

Assuming that all of the characteristics are step-wise function and applying the classical method of transformation of random variables [15] the probability density function of the quantizers' output signal η can be derived as a weighted sum of Dirac delta-functions [8] The obtained expressions were applied to find the output variance $\sigma_y^2 = \mathbb{D}\{\eta\}$, usually related to output process' power.

For the adopted models (*A* to *F*) $\mathbb{D}\{\eta\}$ is given by

$$\mathbb{D}\left\{\eta_{dir_1}\right\} = \left(1-2^{-R}\right)^2\left[1+2\Phi\left(-1+2^{-R}\right)\right] + \sum_{\substack{m=-2^R+1 \\ m\neq\{-1;0\}}}^{2^R-2} 2^{-2R}m^2\left[\Phi\left(2^{-R}\left(m+1\right)\right)-\Phi\left(2^{-R}m\right)\right], \quad (7)$$

$$\mathbb{D}\left\{\eta_{dir_2}\right\} = \sum_{\substack{m=-2^R+1 \\ m\neq\{-1;0\}}}^{2^R-2} 2^{-2R}m^2\left[\Phi\left(2^{-R}\left(m+1\right)\right)-\Phi\left(2^{-R}m\right)\right], \quad (8)$$

$$\mathbb{D}\left\{\eta_{dir_3}\right\} = \sum_{\substack{m=-\infty \\ m\neq\{-1;0\}}}^{\infty} 2^{-2R}m^2\left[\Phi\left(2^{-R}\left(m+1\right)\right)-\Phi\left(2^{-R}m\right)\right], \quad (9)$$

$$\mathbb{D}\left\{\eta_{comp_1}\right\} = \left(1-2^{-R}\right)^2\left[1-\Phi\left(1-2^{-R}\right)\right]+\Phi\left(-1+2^{-R}\right) + \sum_{\substack{m=-2^R+1 \\ m\neq 0}}^{2^R-2} 2^{-2R}m^2\left[\Phi\left(2^{-R}\left(m+1\right)\right)-\Phi\left(2^{-R}m\right)\right], \quad (10)$$

$$\mathbb{D}\left\{\eta_{comp_2}\right\} = \sum_{\substack{m=-2^R+1 \\ m\neq 0}}^{2^R-2} 2^{-2R}m^2\left[\Phi\left(2^{-R}\left(m+1\right)\right)-\Phi\left(2^{-R}m\right)\right], \quad (11)$$

$$\mathbb{D}\left\{\eta_{comp_3}\right\} = \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} 2^{-2R}m^2\left[\Phi\left(2^{-R}\left(m+1\right)\right)-\Phi\left(2^{-R}m\right)\right]. \quad (12)$$

In all of the above expressions $\Phi(\bullet)$ stands for the cumulative probability density function of the standard normal process.

One can see that the difference between (7-9) and (10-12) is mainly due to asymmetry (respective to the vertical axis) of the quantization characteristic for the complementary code. Hence the discrepancy between the variances of the two forms of number representation is because of different excluding points.

It should be pointed out that the prevailing ADCs' description assumes that the only source of the power distortion in a quantizer is the classical quantization noise η_{cqn} that is uniformly distributed and depends on the bit depth only (and not the input signal power), with the variance , described as $\sigma_q^2 = \mathbb{D}\left\{\eta_{cqn}\right\} = 2^{-2R}/3$ for both adopted cases of number representation.

IV. Simulation and Analysis

To demonstrate the applicability of the derived results the input-output power relation was analyzed for all of the assumed cases (see Fig. 1-2 for models *A* to *C* and Fig. 3-4 for models *D* to *F*). In order to comply with the idealized one-to-one mapping a bisector is plotted on each figure. The black dashed line of Fig. 1-4 indicates the variance for the classical quantization noise model (in Fig. 2 and 4 it not visible since the adopted plotted region, i.e. σ_q^2 is much smaller , hence it is out of the bounds).

The performed analysis made it possible to conclude that for the case of the direct code with truncation number representation the quantizer suppresses the inputs signal power. And the less the bit depth, the greater is the suppression level.

It should be noted that the physics behind the effect is different for the cases of large and small input signals: the small ones almost completely fall into the zero quantization step (round the origin) at the same time powerful signals (comparing to the unit aperture) put the ADC into overload, hence the large deviations of the instantaneous values are being either saturated or nullified. Moreover, one can notice that the reset strategy exhibits much stronger power suppression.

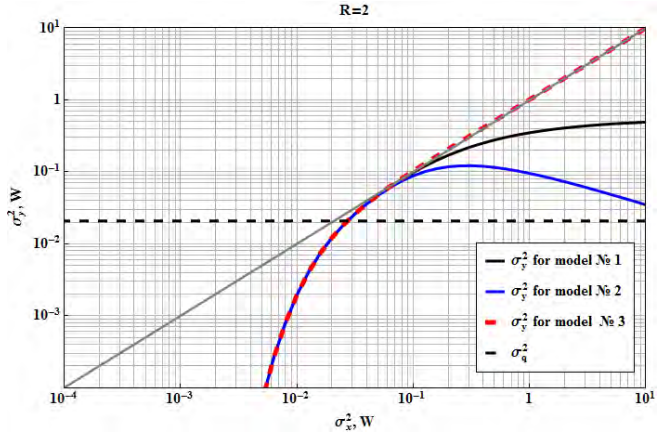


Fig. 1. Output signal power for the case of direct code representation with *R*=2

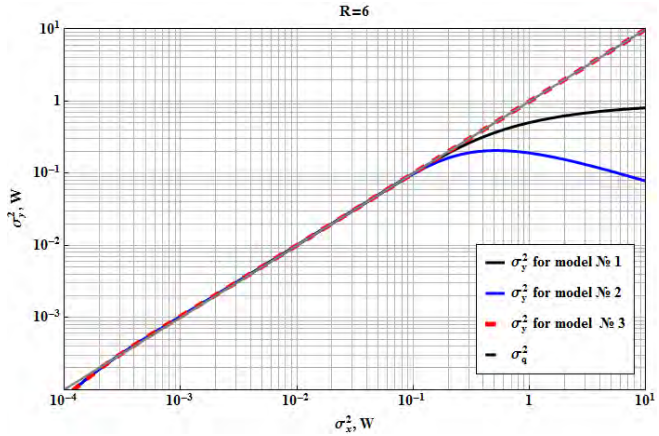


Fig. 2. Output signal power for the case of direct code representation with *R*=6

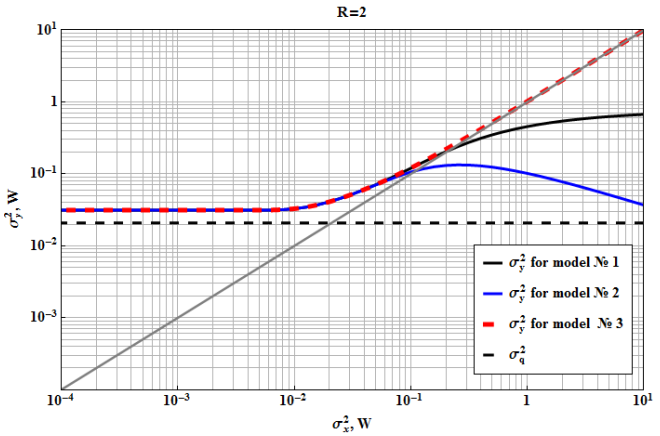


Fig. 3. Output signal power for the case of complementary code representation with *R*=2

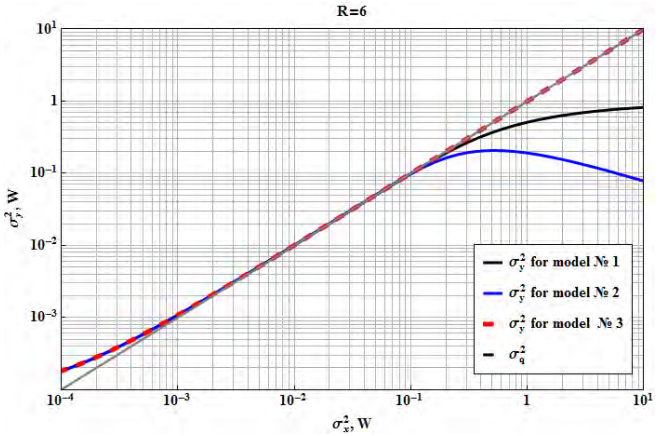


Fig. 4. Output signal power for the case of complementary code representation with *R*=6

It can be seen that (compare Fig. 1 and 2 or Fig. 3 and 4) the increase in bit depth helps to better reproduce smaller signals but does not improve the situation for large ones. It was also noted that starting from *R*=6 the difference in the results is almost negligible, hence it can be assumed as a reference for bit depth choice in practical applications.

Furthermore, for a two-bit quantizer the optimal input signal power, which causes minimum power suppression is around 0.1 W. This point can be used as a reference dividing the signals into small and large since the idealized quantizer (without overload and infinite characteristic) before this it behaves like all other models (with overload) and after it like ideal one-to-one mapping device.

Contrary to the direct code for complementary number representation the small input signals cause the increase in the output power. With this exception all other effects, mentioned above, hold true.

In order to perform a quantitative rather than a qualitative description, a power distortion coefficient was defined as $k = \sigma_y^2 / \sigma_x^2$.

Fig. 5-8 present *k* (scaled logarithmically) for all of the assumed models with *R*=2 and *R*=6.

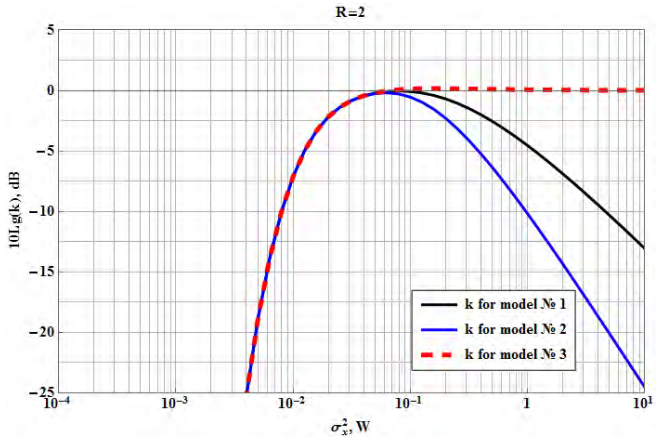


Fig. 5. Power distortion coefficient for the case of direct code representation with R=2

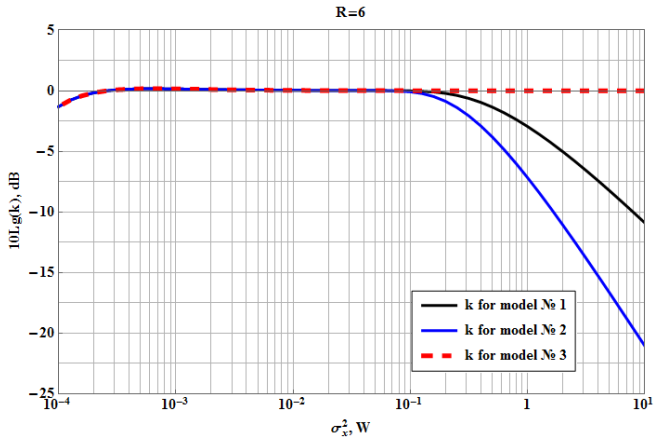


Fig. 6. Power distortion coefficient for the case of complementary code representation with R=6

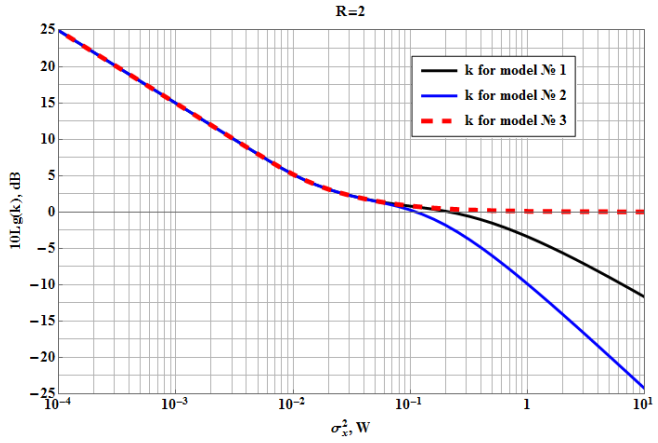


Fig. 7. Power distortion coefficient for the case of complementary code representation with R=2

It can be observed that the increase of bit depth leads to the increase in the range of possible input signal powers that case minimal distortion. For example, adopting as an admissible level ±3 dB referenced to the minimum distortion one can see that the input range is:

- for direct code (R=2): from 0.15 mW to 0.25 W (for reset)/ 0.7 W (for saturation);

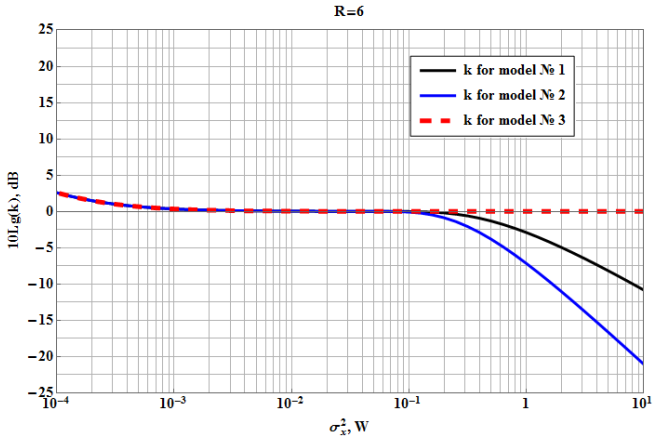


Fig. 8. Power distortion coefficient for the case of complementary code representation with R=6

- for direct code (R=6): greater than 0.4 W (for reset)/ 1.2 W (for saturation);
- for complementary code (R=2): from 0.15 mW to 0.25 W (for reset)/ 0.9 W (for saturation);
- for complementary code (R=6): greater than 0.4 W (for reset)/ 1.3 W (for saturation).

The increase of the input signal (more than the reference point) as large as 10 W leads to the output signal power suppression up to 21 dB (for both number representation codes, saturation and R=2) and 11 dB (for both number representation codes, saturation and R=6).

V. Conclusion

The performed research assumes the problem of the statistical description of the ADC's output signal power distortion. The main cause of such an effect is the essential nonlinearity of the deployed quantizer. The results of the research are obtained under the assumption that the quantizer is driven by the input zero mean circular symmetric Gaussian process, which in many practical situations reflects the statistics of the input signal, being the samples of complex signal-code constructs. An analysis of the power suppression induced by the ADC was performed depending on the bit resolution for several specific practical situations: two number representation forms (direct and complementary codes with truncation) and two ways of overloading resolving (saturation and reset). The results are quantitatively compared with the idealized quantization device exhibiting no overloading, hence infinite characteristic and classic linear quantization model with the uniform error distribution. The possible ranges of analyzed parameters values, which yield minimum suppression, are calculated. On the opposite, parameter values that do not allow by no chance to reach the desired signal transition quality (which constitutes to the worst case) are determined.

VI. References

[1] J. Bard and V. Kovarik, Software defined radio: The Software Communications Architecture, Wiley, 2007

[2] H. Venkataraman and G. Muntean, Cognitive radio and its application for next generation cellular and wireless networks. Dordrecht: Springer, 2012.

[3] "802.22b-2015 - IEEE Standard for Information Technology--Telecommunications and information exchange between systems -

Wireless Regional Area Networks (WRAN)--Specific requirements - Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications:Policies and Procedures for Operation in the TV Bands - Amendment 2: Enhancement for Broadband Services and Monitoring Applications - IEEE Standard", ieeexplore.ieee.org, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/7336461

[4] "5G Radio Access : Capabilities and Technologies", 5g.co.uk, 2020. [Online]. Available: https://5g.co.uk/white-papers/5g-radio-access-capabilities-and-technologies.

[5] A. Paulraj, D. Gore and R. Nabar, Introduction to space-time wireless communications. Cambridge: Cambridge Univ. Press, 2008.

[6] F. Boccardi, R. Heath, A. Lozano, T. Marzetta and P. Popovski, "Five disruptive technology directions for 5G", IEEE Communications Magazine, vol. 52, no. 2, pp. 74-80, 2014. DOI: 10.1109/mcom.2014.6736746

[7] C. Rodenbeck, M. Martinez, J. Beun, J. Silva-Martinez, A. Karsilayan and R. Liechty, "When Less Is More … Few Bit ADCs in RF Systems", IEEE Access, vol. 7, pp. 12035-12046, 2019. DOI: 10.1109/access.2018.2890701

[8] B. Widrow and I. Kollar, Quantization noise. Cambridge: Cambridge University Press, 2008.

[9] Yu.A. Bryukhanov, "A method of analysis of periodic processes in nonautonomous discrete-time systems with quantization", Journal of Communications Technology and Electronics, vol. 53, no. 7, pp. 807-813, 2008.

[10] J. Proakis and M. Salehi, Digital communications. Boston: McGraw Hill, 2008

[11] S. Wei, D. Goeckel and P. Kelly, "Convergence of the Complex Envelope of Bandlimited OFDM Signals", IEEE Transactions on Information Theory, vol. 56, no. 10, pp. 4893-4904, 2010. DOI: 10.1109/tit.2010.2059550

[12] H. Yoo, F. Guilloud and R. Pyndiah, "Probability distribution analysis of M-QAM-modulated OFDM symbol and reconstruction of distorted data", EURASIP Journal on Advances in Signal Processing, vol. 2011, no. 1, 2011. DOI: 10.1186/1687-6180-2011-135

[13] Z. Qu, I. Djordjevic and J. Anderson, "Two-Dimensional Constellation Shaping in Fiber-Optic Communications", Applied Sciences, vol. 9, no. 9, p. 1889, 2019. DOI: 10.3390/app9091889

[14] W. Jung, Op Amp applications handbook. Amsterdam: Newnes, 2005.

[15] S. D. Kulchycki, “Continuous-Time Sigma-Delta ADCs”, National Semiconductor, 2008.

Algorithm of Generalized Solution an Optimal Control Problems for First-Order Differential Equations with Riemann-Hilbert Boundary Conditions

David Devadze
Department of Computer Sciences
Batumi Shota Rustavely State University
Batumi, Georgia
david.devadze@gmail.com
david.devadze@bsu.edu.ge

Vakhtang Beridze
Department of Computer Sciences
Batumi Shota Rustavely State University
Batumi, Georgia
vakhtangi@yahoo.com
v.beridze@bsu.edu.ge

***Abstract* - The present paper is devoted to optimal control problems whose behavior is described by first-order differential equations on the plane with Riemann-Hilbert Boundary conditions. Necessary and sufficient optimality conditions are obtained. A numerical algorithm for the solution of an optimal control problem for generalized analytic functions is given.**

Keywords — Optimal control problems, Riemann-Hilbert boundary conditions

I. INTRODUCTION

For many optimization problems in elasticity theory, mechanics, diffusion processes, the kinetics of chemical reactions and so on, the state of a system is described by partial differential equations. Therefore, optimal control problems for systems with distributed parameters attract much attention. [1].

When dealing with questions of optimization for systems with distributed parameters, an important tool is the use of existence problems for generalized solution under discontinuous right-hand parts of the equation. In the previous work [2], a unified scheme is formulated for proving necessary conditions of optimality for a wide class of object optimization problems with distributed parameters.

Necessary optimality conditions are established by using the approach worked out in [3]-[4] for controlled systems of general type.

The paper is devoted to optimal control problems whose behavior is described by linear first-order differential equations on the plane with Riemann-Hilbert Boundary conditions [5]-[6].

A numerical algorithm [7]-[8] of the optimal control problem solution for generalized analytic functions is given.

II. NECESSARY AND SUFFICIENT OPTIMALITY CONDITIONS

Let G be a bounded domain of the plane with boundary Γ, $z = x + iy \in G$, $\partial_{\bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)$ is the generalized Sobolev derivative [9]. Let U be some bounded subset from E. Each function $u(z): G \to U$ will be called a control. The set U is called the control domain. We call the function $u(z)$ an admissible control if $u(z) \in L_p(G)$, $p > 2$. The set of all admissible controls is denoted by Ω.

For each fixed $u \in \Omega$ consider the following boundary value problem:

$$\partial_{\bar{z}} w = A_1(z)w + A_2(z)\bar{w} + A_3(z)u + A_4(z), \quad z \in G, \qquad (1)$$

$$\mathrm{Re}[\overline{\lambda(z)}\, w(z)] = g(z), \quad z \in \Gamma. \qquad (2)$$

We will assume, that $A_i \in Lp(G)$, $p > 2$, $i = \overline{1,4}$, $\lambda, g \in c_\mu(\bar{G})$, $|\lambda(z)| = 1$.

If the task index (1) - (2) $n \geq 0$, then in this case, the boundary conditions (2) attach the following normalization condition

$$\mathrm{Im}[\overline{\lambda(z_j)}\, w(z_j)] = c_j, \qquad j = \overline{1, 2n+1} \qquad (3)$$

If the index of the boundary value problem $n < 0$, then these normalization conditions are redundant, since in this case the task (1) - (2) has the only solution, if the solvability conditions are implemented [4].

Consider the functional

$$I(u) = \mathrm{Re}\iint_G [B_1(z)\, w + B_2(z)\, u]\, dxdy \qquad (4)$$

and system of restrictions

$$L_k(u) = \mathrm{Re}\iint_G [c_{1k}(z)w + c_{2k}(z)u]dxdy, \qquad k = \overline{1,r}, \qquad (5)$$

where functions $B_i, C_{ik} \in Lp(G)$, $p > 2$, $i = 1,2$, $k = \overline{1,r}$, r - arbitrary natural number.

We pose the following problem: Find function $u_0 \in \Omega$, in which the solution of the Riemann-Hilbert boundary value problem (1) - (3) satisfies the restrictions system (5) and gives to functional (4) minimal value.

For problem (1) - (5), the following theorem is proved by evaluating the first variations according to the general scheme of the Pontryagin maximum principle [5]-[10].

Theorem. Let P cosine of R^{r+1} and functions $\psi_0, \psi_{0k} (k = \overline{1,r})$ are respectively solutions of the following equations:

$$\begin{cases} \partial_{\bar z}\psi + A_1(z)\psi + \overline{A_2(z)}\bar\psi + B_1(z) = 0, \ \ z \in G, \\ \mathrm{Re}[\lambda(z)\psi(z) = 0, \ z \in \Gamma, \end{cases} \tag{6}$$

$$\begin{cases} \partial\bar z\psi_k + A_1(z)\psi_k + \overline{A_2(z)}\bar\psi_k + c_{1k}(z) = 0, \quad z \in G, \\ \mathrm{Re}[\lambda(z)\psi_k(z) = 0, \quad z \in \Gamma, k = \overline{1,2}. \end{cases} \tag{7}$$

Then for any $\eta \in P$ almost on G the fulfillment of the following relation

$$\begin{aligned} &\mu_0\,\mathrm{Re}[(A_3(z)\psi_0(z) + B_2(z))u_0(z)] + \\ &+ \mathrm{Re}[\sum_{k=1}^{r}\mu_k(A_3(z)\psi_{0k}(z) + c_{rk}(z))u_0(z)] = \\ &= \inf_{u\in U}\{\mu_0\,\mathrm{Re}[(A_3(z)\psi_0(z))u] + \\ &+ \mathrm{Re}[\sum_{k=1}^{r}\mu_k(A_3(z)\psi_{0k}(z) + c_{rk}(z))u]\} \end{aligned} \tag{8}$$

is necessary and sufficient condition for the optimality of $(u_0 w_0)$.

III. A NUMERICAL METHOD FOR SOLVING THE OPTIMAL CONTROL PROBLEM FOR GENERALIZED ANALYTIC FUNCTIONS

Note, that the solution (u_0, w_0) of the optimal control problem (1) - (5) in case of the condition of general position [10], by the above theorem is possible to find following in sequence: solve (6), (7) equations, then from the relation (8) determine the optimal u_0 and consequently from (1) - (3) - optimal solution w_0.

Therefore, to construct a numerical method for solving the optimal control problem (1) - (5), it is enough to investigate the numerical method for solving the following problem:

$$\partial_{\bar z}\psi + A(z)\bar\psi + F(z) = 0, \ \ z \in G, \tag{9}$$

$$\mathrm{Re}[\bar\gamma(s)\psi(z(s)) = c(s), \tag{10}$$

Where $\psi = u + iv$, $z = x + iy$, G - simply connected bounded domain, A, F - known complex functions of Lp(G), $\gamma(s) = \alpha(s) + i\beta(s)$, s - arc length on Γ, α, β, c - functions from $C_\mu(\Gamma)$, $0 < \mu < 1$, $|\gamma(s)| = 1$.

Consider the case, when the index of the problem $n \geq 0$. As already noted in [5], under additional normalization conditions (3) problem (9), (10) has the only solution from $C_\mu(\bar G)$.

To find an approximate solution to the problem is more convenient with additional conditions which highlight the single solution, define integral type functionals [11]:

$$\int_\Gamma \mathrm{Im}[\gamma\psi]h_\alpha(s)ds = C_\alpha, \qquad \alpha = \overline{0,2n}, \tag{11}$$

where C_k - arbitrary real constants, $k = \overline{0,2n}$, and functions $h_\alpha(s)$ constitute areal, continuous, periodic system on Γ function, which has property, that for any 2μ points on Γ, $\mu \leq n$, exists linear combination $\sum_{\alpha=0}^{2n}\lambda_\alpha h_\alpha(s)$ (λ_α - valid constants), changing sign in these and only these 2μ points. Such functions may be, for example,

$$1,\ \cos\left(\frac{2\pi}{D}s\right),\ \sin\left(\frac{2\pi}{D}s\right),\ \ldots,\ \cos\left(\frac{2\pi}{D}ns\right),\ \sin\left(\frac{2\pi}{D}ns\right),$$

Where D - curve length Γ.

Let's prove, that the problem (9) - (11) is posed correctly. By correctness we mean the unique solvability of the problem and the continuous dependence of the solution on the right side (9) - (11), defined by inequality.

$$\|\psi\|_{C_\mu(\bar G)} \leq c\left[\|F\|_{Lp(G)} + \|c(s)\|_{C\mu(\bar G)} + \sum_{\alpha=0}^{2n}c_\alpha^2\right]. \tag{12}$$

A similar estimate in the norm of space $L_2(G)$ is received in [11].

To prove the inequality (12) reduce the problem (9) - (11) to an integral equation with a completely continuous operator. For this, we introduce the operator

$$T_G[\psi] = -\frac{1}{\pi}\iint_G \frac{\psi(t)}{t-z}d\xi d\eta, \quad t = \xi + i\mu, \tag{13}$$

And operator $S_G[\psi]$, which displays $L_p(\bar G)$ in a subset of analytical functions, which satisfies the following conditions at the border Γ:

$$\begin{aligned} &\mathrm{Re}[\bar\gamma(T_G[\psi] + S_G[\psi])] = 0, \quad z \in \Gamma, \\ &\int_\Gamma \mathrm{Im}[\bar\gamma(T_G[\psi] + S_G[\psi])]h_\alpha(s)\,ds = 0, \quad \alpha = \overline{0,2n}. \end{aligned} \tag{14}$$

As mentioned above, these conditions $S_G[\psi]$ defined ambiguously.

Define the operators

$$\begin{aligned} &P(\psi) = T_G[\psi] + S_G[\psi], \\ &P_A(\psi) = P(A\bar\psi). \end{aligned} \tag{15}$$

We introduce the function $\varphi(z)$, which is an analytic function z in area of G and meets the following conditions on Γ:

$$\begin{aligned} &\mathrm{Re}[\bar\gamma\varphi(z)] = c(s), \quad s \in \Gamma, \\ &\int_\Gamma \mathrm{Im}[\bar\gamma\varphi]h_\alpha(s)ds = 0, \quad \alpha = \overline{0,2n}. \end{aligned}$$

If we now consider, that $\partial_{\bar z}P(\psi) = \psi(z)$ [9], it is easy to see that the solution to the problem (9) - (11) is the solution of the following functional equation:

$$\psi(z) = P_A(\psi) + P(F) + \varphi(z) + \sum_{\alpha=0}^{2n}c_\alpha\psi_\alpha(z), \tag{16}$$

where $\psi_\alpha(Z)$ - analytic function system, satisfying on Γ in condition $\mathrm{Re}[\bar\gamma\psi_\alpha] = 0$, $\alpha = \overline{0,2n}$, and conditions

$\int_\Gamma \mathrm{Im}(\bar{\gamma}\psi_\alpha)h_k(s)ds = \delta_{\alpha k}$, $k = \overline{0,2n}$, $\alpha = \overline{0,2n}$, $\delta_{\alpha\kappa}$ Kronecker symbol.

It is clear, that the problem (9) - (11) and (16) are equivalent, so, any solution of the first problem is the solution to the second problem and, conversely [2]-[13]. To prove (12) is considered tasks (16). Using the well-known properties of operators T_G and P [10] can show, that they are completely continuous operators over the field of real numbers. Obviously, the operator $P_A(\psi) = P\left(A\bar{\psi}\right)$ also completely continuous.

Considering homogeneous tasks that match the tasks (9) - (11) and (16) ($F(z) = 0$, $c(s) = 0$, $c_\alpha = 0$, $\alpha = \overline{0,2n}$), we can conclude that these problems have only trivial solutions. From the complete continuity of the operator $P_A(\psi)$ last remark follows the existence and boundedness of the operator $(I - P_A)^{-1}$ ($I -$ current operator).

Let denote $\left\|(I - P_A)^{-1}\right\| = M$, $\|p\| = M_\rho$, where M and M_p - are permanent. Considering a priori estimates [2]

$$\begin{aligned}
&\left\|\sum_{\alpha=0}^{2n} C_\alpha \psi_\alpha(z)\right\|^2 \le M_0^2 \sum_{\alpha=0}^{2n} c_\alpha^2, \quad M_0 = \mathrm{const},\\
&\left\|\sum_{\alpha=0}^{2n} C_\alpha \psi_\alpha(z)\right\|^2 \le M_0^2 \sum_{\alpha=0}^{2n} c_\alpha^2, \quad M_\varphi = \mathrm{const},\\
&\|P(F)\|_{L_p(G),C_\mu(\bar{G})} \le M_P \bullet \|F\|_{L_p(G)},
\end{aligned} \tag{17}$$

from (16) we obtain

$$\|\psi\|_{C_\mu(\bar{G})} \le M \cdot \max(M_0, M_\varphi, M_p)\left[\|F\|_{L_p(G)} + \|C\|_{C_\mu(\bar{G})} + \sum_{\alpha=0}^{2n} C_\alpha^2\right],$$

which meets (12) under $C = M \bullet \max(M_0, M_\varphi, M_p)$. Hence, the correctness of the problem (9) - (11) is proven.

All arguments for the correctness of the problem (9) - (11) turn out quite plain for the occasion, when G represents a circle of unit radius: $|z| = 1$ и $\gamma = z^{-n}$. in this case, the equation (16) will look [9].

$$\begin{aligned}
\psi(z) = {} & \frac{-1}{\pi}\iint_G \left[\frac{A\bar{\psi}}{t-z} + \frac{z^{2n+1}A\bar{\psi}}{1-\bar{t}z}\right]d\xi d\eta - \\
& - \frac{1}{\pi}\iint_G \left[\frac{F}{t-z} + \frac{z^{2n+1}\bar{F}}{t-z}\right]d_\xi d_\eta + \frac{z^n}{2\pi i}\int_\Gamma c(s)\frac{t+z}{t-z}\frac{dt}{t} + \\
& \sum_{k=0}^{n-1}\left\{\frac{1}{2\pi}\int_\Gamma c(s)\cos(n-k)sds(z^k - z^{2n-k}) + \right. \\
& \left. + \frac{i}{2\pi}\int_\Gamma c(s)\sin(n-k)sds(z^k + z^{2n-k})\right\} + \\
& + \sum_{k=0}^{n-1}\left[\alpha_k(z^k - z^{2n-k}) - i\beta_k(z^k + z^{2n-k})\right] + i\beta_n z^n,
\end{aligned} \tag{18}$$

α_k и β_k - permanent, $t = \xi + i\eta$. Here

$$P(\psi) = \frac{-1}{\pi}\iint_G \left[\frac{\psi(t)}{t-z} + \frac{z^{2n+1}\bar{\psi}}{1-\bar{t}z}\right]d\xi d\eta$$

$$\varphi(z) = \frac{z^n}{2\pi i}\int_\Gamma c(s)\frac{t+z}{t-z}\frac{dt}{t} +$$

$$+ \sum_{k=0}^{n-1}\left\{\frac{1}{2\pi}\int_\Gamma c(s)\cos(n-k)sds(z^k - z^{2n-k}) + \right.$$

$$\left. \frac{i}{2\pi}\int_\Gamma c(s)\sin(n-k)sds(z^k + z^{2n-k})\right\}$$

System of function $h_\alpha(s)$ has the form 1, cos(s), sin(s), cos(ns), sin(ns). For those equations is obvious, that operator $P(\psi)$ will completely continuous as an operator with a weak singularity, the estimate for the norm is obtained immediately from the expression $\varphi(z)$ and, hence, evaluation (17) and (12) becomes apparent.

In case of $(m+1)$ - connected area G, for solving problem (9) - (11) evaluation (12) in the norm of space $L_2(G)$ has the form [11]

$$\iint_G |\psi|^2 dxdy \le C\left[\iint_G |F|^2 dxdy + \int_\Gamma c^2(s)ds + \sum_{\alpha=0}^{2n-m} C_\alpha^2\right] \tag{19}$$

note that in the case of a circle from (18) immediately evaluate (19).

To compose a difference system of equations, we replace the problem (9) - (11), with method of least squares, equivalent of variation problem. Consider the problem of finding the minimum of functional

$$I(\psi) = \iint_G \left|\partial_{\bar{z}}\psi - A\bar{\psi} - F\right|^2 dxdy + \int_\Gamma [\mathrm{Re}(\bar{\gamma}\psi) - c(s)]^2 ds \tag{20}$$

in class of function $C_\eta(\bar{G})$, satisfying conditions (3). We will solve the variation problem by the grid method [11]. Cover the area G with grid of isosceles right-angled triangles with the cathetus h. Received mesh area is denoted by G_h. Let $\bar{G}_c \subset G_h$ and $\bar{G}h$ is minimum triangular cover region G. Many corner points $\bar{G}_h$ also denoted by $\bar{G}_h$

Consider the problem of finding the minimum functional (20), under conditions (11) in class of function $\hat{\psi}_h$:

$$\hat{\psi}_h(x,y) = \hat{\psi}_{ij} + \frac{\hat{\psi}_{ij+1}\hat{\psi}_{ij}}{h}(y - y_j) + \frac{\hat{\psi}_{i+1,j-}\hat{\psi}_{ij}}{h}(x - x_i), \tag{21}$$

defined and continuous in the field $\bar{G}_h$ and linear in each triangle $\Delta_{pq.}$ in field of $\bar{G}h$. Here Δ_{pq} - triangle with vertices at points (pq), (p+1,q), (p,q+1), $\hat{\psi}_{ij}$ - meaning $\hat{\psi}_h$ in nodes if grid. A function that minimizes functional (20) in class of function $\hat{\psi}_h$, denote by ψ_h, and its value in the nodes of the grid - through ψ_{ij}. through ψ_h we also denote the vector with components ψ_{ij}, Γ_h - grid points of $\bar{G}_h$,

which are the vertices of the triangles and intersect with Γ, $G_h \equiv \bar{G}_h / \Gamma_h$.

For points of grid $(p,q) \in \bar{G}_h$ we have $(\hat{\psi}_h)_{pq} = \hat{u}_{pq} + i\hat{\vartheta}_{pq}$ and $\psi_{pq} = u_{pq} + i\vartheta_{pq}$. Arrange the values $\hat{u}_{pq}$ and $\hat{u}_{pq}$ in a certain sequence $\hat{w}_k (k = 1,2,....,N)$, where N equal to double the number of points $\bar{G}_h$. Will assume, that the boundary points Γ_h match the first k values $\hat{w}_1, \hat{w}_2, ..., \hat{w}_k$. Similarly, we build $w_1, w_2, ..., w_k, ...w_N$. Then the expression of the functional (20) and conditions (11) on element $\hat{\psi}_h$, certain formula (21), will take the following look:

$$I(\hat{\psi}_h) = I(\hat{w}_1, \hat{w}_2, ..., \hat{w}_N) = \sum_{i,j=1}^{N} u_{ij}\hat{w}_j\hat{w}_j + 2\sum_{i=1}^{N} b_i\hat{w}_i + d, \quad (22)$$

$$\sum_{j=1}^{k} c_{\alpha j}\hat{w}_j = c_\alpha, \quad \alpha = 0,1,...2n. \quad (23)$$

Where c_α - given numbers, a_{ij}, b_i, d, $c_{\alpha j}$ can be calculated from (20). Minimum of functional (22) under conditions (23) gives an approximate solution to the problem (9) - (11).

For Lagrange functions (λ_α - Lagrange multipliers)

$$\hat{I}(\hat{w}_1, \hat{w}_2, ..., \hat{w}_N) = I(\hat{w}_1, \hat{w}_2, ..., \hat{w}_N) + \\ +2\sum_{\alpha=0}^{2n}\lambda_\alpha\left(\sum_{j=1}^{k} c_{\alpha j}\hat{w}_j - c_\alpha\right), \quad (24)$$

if we write the minimum condition, we obtain a system of equations for determining $\psi_h = \{w_1, w_2, ..., w_N\}$:

$$\sum_{j=1}^{N} a_{ij}w_j + \sum_{\alpha=0}^{2n}\lambda_\alpha c_{\alpha i} + b_i = 0, \quad i = 1,2...,k; \quad (25)$$

$$\sum_{i=1}^{N} a_{ij}w_j + b_i = 0, \quad i = k+1,...,N; \quad (26)$$

$$\sum_{j=1}^{k} c_{\alpha j}w_j = c_\alpha, \quad \alpha = 0,2,...,2_n. \quad (27)$$

Following [11] will show, that system (25) - (27) in solvable for any b_i, c_α. For this is enough to show, that homogeneous system (25) - (27) ($b_i = 0$, $j = \overline{1,N}$, $c_\alpha = 0$, $\alpha = \overline{0,2n}$) has only zero solution.

Let $w_1^0, w_2^0, ..., N_N^0$ - is homogeneous system solution, then by multiplying the equations (25) - (26), respectively, on $w_1^0, w_2^0, ..., w_N^0$, and after sum them up, we get

$$0 = I(w_1^0, w_2^0, ..., w_N^0) + \sum_{\alpha=0}^{2n}\lambda_\alpha\sum_{j=1}^{k} c_{\alpha j}w_j^0 = \\ = \iint_G \left|\partial_{\bar{z}}\psi_2^0 - A\bar{\psi}_h^0\right| dxdy + \int_\Gamma [Re(\bar{\gamma}\psi_h^0)]^2 ds.$$

From here we get

$$\partial_{\bar{z}}\psi_h^0 = A\bar{\psi}_h^0, \quad z \in G,$$

$$Re[\bar{\gamma}\psi_h^0] = 0, \quad z \in \Gamma.$$

Therefore, given (27) and (11), under $c_\alpha = 0$, we conclude that $\psi_h^0 \equiv 0$ in $\bar{G}$.

So, the difference solution $\psi_h = \{w_1, ..., w_N\}$ exists for any b_i, c_x and is determined using the system (25) - (27).

Let ψ - is exact solution to the problem (9) - (11). As known [9]:

$$\psi(z) = \psi_0(z) + \sum_{\alpha=1}^{2n+1}\nu_\alpha\psi_\alpha(z), \quad (28)$$

Where ψ_0 - particular solution to the problem (1) - (2), ψ_α, $\alpha = \overline{1,2n+1},-$ linearly independent solutions of the homogeneous problem (9) - (10), ν_α - real constants, which are determined from the conditions (3) as solutions to the following system of algebraic equations:

$$\int_\Gamma Im[\bar{\gamma}\psi_0]h_\ell(s)ds + \\ +\sum_{\alpha=1}^{2n+1}\nu_\alpha\int_\Gamma Im[\bar{\gamma}\psi_\alpha]h_\ell(s)ds = c_\ell, \quad \ell = 0,1,...,2n. \quad (29)$$

Let function ψ_h implements a minimum of functionality (20) under conditions (11). We prove the convergence of the average ψ_h to ψ.

Considering (9) - (11), we have

$$I(\hat{\psi}_h) = \iint_G \left|\partial_{\bar{z}}(\hat{\psi}_h - \psi) - A(\hat{\psi} - \psi)\right|^2 dxdy + \\ +\int [R\,e(\bar{\gamma}_h - \psi))]^2 ds, \quad (30)$$

$$\int_\Gamma Im[\bar{\gamma}(\hat{\psi}_h - \psi)]h_\alpha(s)ds = 0, \quad \alpha = 0,1,...,2n. \quad (31)$$

For a difference solution $\hat{\psi}_h$ we have $I(\psi_h) = \min I(\hat{\psi}_h)$, where minimum is taken for all $\hat{\Psi}_h$, which satisfy the conditions (31).

For any ψ_α, $\alpha = \overline{0,2n+1}$ from (20) define a function $\hat{\psi}_\alpha$, continuous in the field $\bar{G}_h$ function $\tilde{\psi}_\alpha$, continuous area $\bar{G}_{h,}$ flat in every triangle $\triangle pq$ and receiving at the vertices of triangles z_{pq} values $\psi_\alpha(z_{pq})$. Make in the region $\bar{G}_h$ function:

$$\tilde{\psi} = \tilde{\psi}_0 + \sum_{\alpha=1}^{2n+1}\tilde{\nu}_\alpha\tilde{\psi}_\alpha,$$

where $\tilde{\nu}_\alpha$ - is determined from the following system ($\ell = 0,1,...,2n+1$):

$$\int_{\Gamma} \mathrm{Im}[\overline{\gamma}\tilde{\psi}_o]h_\ell(s)ds + \sum_{\alpha=1}^{2n+1} \tilde{\nu}_\alpha \int_{\Gamma} \mathrm{Im}[\overline{\gamma}\tilde{\psi}_\alpha]h_\ell(s)ds = c_\ell .$$

Note that the functions $\tilde{\psi}$ converge uniformly to the function $\psi \in c_\mu(\bar{G})$, when $h \to 0$ in a closed area $\bar{G}_h$. Given the definition of a generalized derivative $\partial_{\bar{z}}$, we conclude, that $\|\partial_{\bar{z}}(\psi - \tilde{\psi})\| L_2(G) \to 0$ at $h \to 0$. Then, from (30) has, that $I(\tilde{\psi}) \to 0$ at $h \to 0$. Hence, $I(\psi_h) \to 0$, $h \to 0$, due to the fact, that $I(\psi_h) \le I(\tilde{\psi})$. From here according (12) and (31) we get

$$\|\psi - \psi_h\|^2_{L_2(G)} \le c\left\{\iint_G \left|\partial_{\bar{Z}}(\psi_h - \psi) - A(\tilde{\psi}_h - \tilde{\psi})\right|^2 dxdy + \right.$$

$$\left. + \int_{\Gamma} [\mathrm{Re}(\overline{\gamma}(\psi_h - \psi))]^2 ds \right\} = c \cdot I(\psi_h)$$

From here we conclude that an approximate solution ψ_h converges to a generalized analytic function ψ in the norm of space $L_2(G)$.

IV. Conclusion

The innovation of the results is following:

1. Are given necessary and sufficient conditions of optimality for linear first-order Differential Equations with Riemann-Hilbert Boundary Conditions.

2. Are constructed numerical algorithms to solve the problems of linear optimal control for generalized analytic functions.

References

[1] A.G. Butkovski, Theory of optimal control of systems with distributed parameters. Nauka, Moscow, 1977.

[2] G.F. Manjavidze and V. Tuchke, "Some boundary-value problems for first-order nonlinear differential systems on the plane," in: Boundary-Value Problems of the Theory of Generalized Analytic Functions and Their Applications, Tbilis. Gos. Univ., Tbilisi (1983), pp. 79–124.

[3] V.I. Plotnikov, Necessary optimality conditions for controllable systems of a general form. (Russian) Dokl. Akad. Nauk SSSR 199 (1971), 275–278.

[4] D.Sh. Devadze and V.Sh. Beridze, Optimality conditions for quasilinear differential equations with nonlocal boundary conditions. Russian Mathematical Surveys. 2013. Vol 68:4 773-775.

[5] D. Devadze, V. Beridze. An Optimal Control Problem for Quasilinear Differential Equations with Bitsadze–Samarski Boundary Conditions. Journal of Mathematical Sciences. April 2015, Volume 206, Issue 4, pp 357–370.

[6] D. Devadze, M. Abashidze, V. Beridze. Solution of an optimal control problem with Mathcad. Proceedings of 2015 IEEE East-West Design and Test Symposium, EWDTS 2015, p.370-375.

[7] M. Abashidze, V. Beridze, D. Devadze. Algorithm of find of the generalized solution of an m-point nonlocal boundary value problem. Proceedings of 2017 IEEE East-West Design and Test Symposium, EWDTS 2017. 2017.

[8] D. Devadze, H. Meladze. Algorithm of Solution an Optimal Control Problem for Elliptic Differential Equations with m-Point Bitsadze-Samarski Conditions. 2018 IEEE East-West Design & Test Symposium (EWDTS), 2018. pp.1-7.

[9] I. N. Vekua, Generalized analytic functions. (Russian) Second edition. Edited and with a preface by O. A. Oleinik and B. V. Shabat. "Nauka", Moscow, 1988.

[10] L.S. Pontryagin, V.G. Boltyanskiĭ, R.V. Gamkrelidze, E.F. Mishchenko, The mathematical theory of optimal processes. (Russian) Fourth edition. "Nauka", Moscow, 1983.

[11] L.S. Klabukova I.I. Chechel. A difference method for solving boundary value problems for generalized analytic functions. USSR Computational Mathematics and Mathematical Physics. Volume 9, Issue 2, 1969, Pages 17-36 .

Optimization Calculation of Thermoelement Linear Dimensions for Microthermoelectric Generator

Vera Loboda,
Roman Buslaev
Peter the Great St. Petersburg Polytechnic University
Russia
vera_loboda@spbstu.ru

***Abstract* — The article deals with the optimization calculation of thermoelement linear dimensions based on the maximization output power criterion of a microthermoelectric generator. The calculations have been made by applying ANSYS Workbench software and genetic algorithm. The thermoelements linear dimensions correlation has been found to remain unchanged when changing thermal boundary conditions. The maximal power dependence on linear dimensions correlation has shown a flat maximum and, the general range change of thermoelements linear dimensions range with a maximum power deviation of ±3 % can be identified.**

***Keywords* — thermoelectric generator, MEMS, finite element method, simulation, optimization, output power.**

I. INTRODUCTION

Solid state thermoelectric generators based on the Seebeck effect convert thermal energy into electricity. Application of microthermoelectric generators (μTEG) to provide reliable and continuous energy supply of low power devices such as wireless sensor networks, smart homes, object monitoring systems and mobile devices have generated an increasing interest recently [1-3]. A μTEG can be integrated into different surfaces of heat sources to transfer energy thus allowing to reduce technical maintenance costs and increase the device service time compared to a battery. It becomes especially critical when a conventional battery is unavailable, or a device is placed either in a distant or aggressive environment.

Thin film manufacturing of TEG has been developing intensely by employing electrochemical methods [4], MEMS [5-7], MBE [8], CVD [9]. A thin film thermoelectric generator is a compact device which has a short period of thermal response and high specific electric power. Silicon application as a base due to its compatibility with CMOS and MEMS processes [10-12], has become an important factor to choose a μTEG.

A typical length of a thermoelectric generator fabricated by using bulk semiconductor technology ranges between 1 mm and 5 mm. The dimension of a thin film thermoelectric generator can be reduced to less than 20 microns [5,6]. When a device is reduced to micron scale, the factors influencing the efficiency of thin-film device must significantly differ from the bulk thermoelectric unit.

This research is targeted at the optimization calculations of thermoelement linear dimensions, including the thermoelement height correlation to the width based on the maximization output power criterion for a microscale thin-film thermoelectric generator.

II. RESEARCH OBJECT AND MODEL DESIGNING

A single thermoelectric generator in its micro scale design has been chosen as a research object. A μTEG consists of two thermoelements (TE) of *n* and *p*-type, TE commutation contact pads area and a substrate. Figure 1 shows a geometric 3D model of a single microthermoelectric generator.

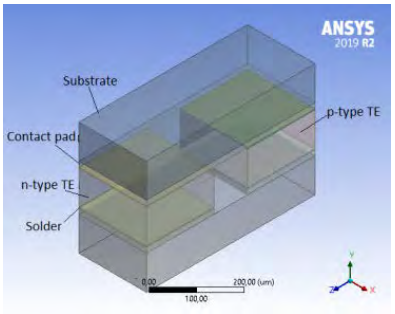


Fig. 1. 3D model of a single microthermoelectric generator

In case there is a temperature gradient between the thermoelectric generator sides, the output power can be calculated as follows [13]:

$$P = U \cdot I = \frac{N^2 \cdot \Delta T^2 \cdot (\alpha_p - \alpha_n)^2 \cdot R_H}{(R_L + R_{in})^2}, \quad (1)$$

where N – thermoelements quantity, ΔT – temperature gradient between thermoelectric generator hot and cold sides; α_n and α_p –the Seebeck coefficients of *n*- and *p*-semiconductor materials, respectively; R_L – load resistivity; R_{in} – generator internal resistivity.

The condition of maximum power transfer to the load is the values equality of the external load and the internal resistivity, that means:

$$P_{max} = \frac{N^2 \cdot \Delta T^2 \cdot (\alpha_p - \alpha_n)^2}{4 \cdot R_{in}} \quad (2)$$

The generator internal resistivity is made of the resistivities of its constituent parts such as thermal elements, contact pads areas and solder:

$$R_{in} = \sum_{i=1}^{N} (R_n + R_p) + \sum_{i=1}^{2N} R_{cont} + \sum_{i=1}^{2N} R_s, \quad (3)$$

where R_n and R_p – resistivities of *n* and *p* thermoelements; R_{cont} – contact pad resistivity; R_s – solder resistivity.

The resistivities of thermoelectric generator constituent parts are calculated by means of materials specific resistivities and linear dimensions of such parts:

$$R_{n,p} = \rho_{n,p} \cdot \frac{H}{L^2}, \quad (4)$$

where ρ_n, ρ_p – materials specific resistivity of *n*- and *p*-thermal elements, respectively; H – thermoelements height; L – square base width of thermoelements.

$$R_{cont} = \rho_{cont} \cdot \frac{h}{L^2}, \quad (5)$$

where ρ_{cont} – material specific resistivity of the contact pad; h – contact pad height; L – square base width of the contact pad.

$$R_s = \rho_s \cdot \frac{h_1}{L^2}, \quad (6)$$

where ρ_s – specific resistivity of solder material; h_1 – solder layer height; L – square base width of the solder layer.

Applying formulas (4) – (6) to formula (3), then applying formula (3) to formula (2), we see as follows:

$$P_{max} = \frac{1}{4} \frac{N \cdot \Delta T^2 \cdot (\alpha_p - \alpha_n)^2}{(\rho_p - \rho_n)\frac{H}{L^2} + 2\rho_{cont} \cdot \frac{h}{L^2} + 2\rho_s \cdot \frac{h_1}{L^2}}. \quad (7)$$

Thus, according to formula (7), it can be stated that the generator internal resistivity and, consequently, the maximal output power depend on the linear dimensions of thermoelectric generator constituent parts, especially thermoelements, due to the significant values difference between specific resistivities of semiconductor and metal materials.

The problem solution to define thermal elements linear dimensions influences on the output power has been carried out based on the finite element method by applying ANSYS Workbench software. The simulation methodology has been described in detail in research articles [14-18]. The important characteristic of this simulation has become the application of DesignXplorer to solve single and multicriteria parametric optimization based on the experiment planning, response surface designing, and correlation analysis and probability evaluation of output parameters deviation from the assigned values.

III. Optimization Calculations

Two series of calculations (preliminary and final) by employing the multi-objective genetic algorithm (MOGA) for different width values of thermal elements (the square base side, L) 200, 100 and 20 µm, and the thermal element height (H) ranged within the limits between 10 and 1000 µm, the contact areas height (h) ranged between 1 and 100 microns. The solder layer thickness and wafer thickness remained constant 3 and 100 µm, respectively, during the calculations.

Bismuth tellurides and antimony of *n*- and *p* type conductivity have been used as functional materials for µTEG thermal elements to provide maximal thermoelectric efficiency in the required temperature range [19]. The simulation initial data include the following physical parameters: the Seebeck coefficient, specific electric resistivity and specific thermal conductivity. These parameters depend significantly on the fabrication technology and have a wide range. So, the evaluation parameters results obtained by the least-squares method were used as the initial parameters [14]. The substrate material and contact pad parameters were selected from ANSYS library and the solder parameters were taken from [20]. Table 1 shows the material physical parameters values of μTEG constituent parts used during the simulation in the temperature range of 300-400 K.

TABLE I. MATERIALS' PHYSICAL PARAMETERS OF MICROTHERMOELECTRIC GENERATOR PARTS

T, K	Zeebeck Coefficient, µV*K^{-1}	Specific Resistivity, µΩ·m	Specific Thermal Conductivity, W/m·K
***n*-type semiconductor material (Bi_2Te_3)**			
300	-148.24	12.02	1.03
350	-152.59	13.48	1.11
400	-156.89	15.05	1.18
***p*-type semiconductor material (Sb_2Te_3)**			
300	208.25	13.33	1.07
350	220.21	16.78	1.06
400	219.52	19.74	1.13
Contact Pad (Cu)			
300	3.5	0.017	406
Solder (Sb/Pb)			
300	-	0.4	48
Substrate (Si)			
300	-	-	148

When simulating the finite elements mesh was generated automatically. The mesh is made of cubic isoparametric elements SOLID226.

The temperature boundary conditions were determined by the temperature of the lower silicon wafer T_h and the temperature of the upper silicon wafer T_c. The following temperature values: T_h=318 K, 343 K, 368 K, 393 K, T_c=293 K have been studied. These temperatures correspond to difference ΔT=25, 50, 75 and 100 K. The model considered electric and thermal contact resistivities on the metal-semiconductor border. It should be noted that the parameters have been selected in accordance with [21]. The simulation results are the output power values of the thermoelectric generator on the external load. The simulation has been carried out for three values of the external load, i.e. 300, 600, 1000 mOhm.

The preliminary optimization calculation has been made to identify the impact of contact pad layer width (h) on the output power. The optimization module was assisted for making the design of experiment matrix for the following values of the geometric dimensions: L=200, 100, 20 µm; H=100, 50 µm; h=1÷100 µm. The geometric dimensions were identified according to the proposed microelectronic technology of microgenerator fabrication that is compatible with CMOS and MEMS technologies. As a result of the calculations, it has been stated that the optimal correlation between the contact pad heights and thermoelements for all TE heights to achieve the maximal power made up h=0,15H.

The main optimization calculation has been carried out to identify the TE linear dimensions influence the output

power taking into account the correlation calculated during the preliminary calculation. Therefore, the updated simulation matrix for the following TE geometric dimensions values: L=200, 100, 20 μm; H=1÷1000 μm has been created in the optimization module. The constant correlation between the heights of the TE and the contact pad $h=0.15H$ has been maintained, and this has become the parameter during solving. The main optimization calculation results are shown in Figures 2, and Table 2.

The following conclusions can be made as a result of the optimization calculations: 1) The correlation of thermoelement height to the width of the square base (H/L) when the maximum power is achieved remains unchanged when changing the thermal boundary conditions. 2) The maximum power value increases monotonically when increasing the temperature difference ΔT with the constant correlation H/L irrespective of the external load resistance value. 3) The maximum power value increases monotonically when increasing the generator thermoelement's width of square base. 4) The ratio H/L remains constant when changing the external load and keeping the temperature mode for L=100 и 20 μm and decreases at L=200 microns proving the idea that the internal resistivity of a μTEG with the optimal dimensions will be in the range of 300 and 600 mOhm. 5) The output power dependencies of a μTEG on the TE height correlation to its width on the load show a flat maximum. The values of H/L are within the range of 0.2÷0,7; 0.2÷0,7 and 0,5÷2 for L=200,100,20 μm, respectively with the deviation ±3% from P_{max} value, irrespective of the external load resistance value. Thus, the general range of thermoelement linear dimensions (H/L) change can be identified as follows: 0.5÷0.7.

CONCLUSION

The provided results of the optimization calculations allow to conclude that such an approach to design a microthermoelectric generator with optimal characteristics is feasible. Unlike similar research, for example, [21,22], this research focuses on including a solder layer between the TE and contact pads as well as according for the influence of the electric and thermal contact the metal-semiconductor border, the contact pad height and the preliminary optimization calculation. These model peculiar features assist in both identifying the TE optimal dimensions and internal resistivity value. All the above mentioned allow to design microthermoelectric generators with the assigned operational conditions (power and temperatures) when the load resistivities are known.

REFERENCES

[1] D. Champier "Thermoelectric generators: A review of applications," Energy Conversion and Management, 2017, vol. 140, pp.167–181.

[2] H.T. Nguyen, V.T. Nguyen, O. Takahito "Flexible thermoelectric power generator with Y-type structure using electrochemical deposition process," Applied Energy, 15 January 2018, vol. 210, pp. 467–476.

[3] F. Deng, H. Qiu, J. Chen, L. Wang, B. Wang "Wearable Thermoelectric Power Generators Combined With Flexible Supercapacitor for Low-Power Human Diagnosis Devices," IEEE Transactions on Industrial Electronics, February 2017, vol. 64, no 2, pp. 1477–1485.

[4] M. J. Kim, T.S. Oh "Thermoelectric Thin Film Device of Cross-Plane Configuration Processed by Electrodeposition and Flip-Chip Bonding," Materials Transactions, 2012, vol.53, no 12, pp. 2160–2165.

[5] G.J. Snyder, J.R. Lim, Huang Chen-Kuo, J.-P. Fleurial "Thermoelectric microdevice fabricated by a MEMS-like electrochemical process," Nature Materials, 2003, vol. 2, Aug. 2003, pp.528-531.

[6] A. Korotkov, V. Loboda, E. Bakulin, S. Dzyubanenko "Fabrication and Testing of MEMS Technology Based Thermoelectric Generator," Proc. 7th Electronics System-Integration Technology Conference (ESTC), September 18-21.2018, Dresden.

[7] E. Bakulin, S. Dzyubanenko, S. Konakov, A. Korotkov, V. Loboda, A.Yugai "Thermoelectric Peltier micromodules processed by thin-film technology," Journal of Physics: Conference Series, 5th International School and Conference on Optoelectronics, Photonics, Engineering and Nanostructures «Saint Petersburg OPEN 2018», vol. 1124, 081005, December 2018.

TABLE II. TEG MAXIMUM OUTPUT POWER FOR DIFFERENT LOAD RESISTIVITYS AND THERMAL BOUNDARY CONDITIONS

	ΔT=100 K				***ΔT=75 K***				***ΔT=50 K***				***ΔT=25 K***			
R_L=300 mOhm																
L, μm	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***
200	50	0,25	1815	738	50	0,25	525	397	50	0,25	108	180	50	0,25	6.2	43
100	35	0,35	360	104	35	0,35	10	55,9	35	0,35	2.2	25,4	35	0,35	0.12	6,04
20	20	1	2e-4	0,25	20	1	1e-4	0,14	20	1	0.1e-4	0,06	20	1	0.1e-5	0,02
R_L=600 mOhm																
L, μm	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***
200	65	0,325	842	711	65	0,325	263	397	65	0,325	53	178	65	0,325	2.9	42
100	35	0,35	40	155	35	0,35	12	84	35	0,35	2.	37,8	35	0,35	0.14	9
20	20	1	4e-4	0,49	20	1	1e-4	0,27	20	1	0.2e-4	0,12	20	1	0.1e-5	0,03
R_L=1000 mOhm																
L, μm	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***	***H, μm***	***H/L***	***U, uV***	***P_{max}, μW***
200	65	0,325	396	629	65	0,325	118	343	65	0,325	4.2	153	65	0,325	1.4	37
100	35	0,35	34	185	35	0,35	10	100	35	0,35	1.2	45	35	0,35	0.12	10,8
20	20	1	6e-4	0,8	20	1	2e-4	0,45	20	1	0.4e-4	0,20	0,20	1	0.3e-5	0,05

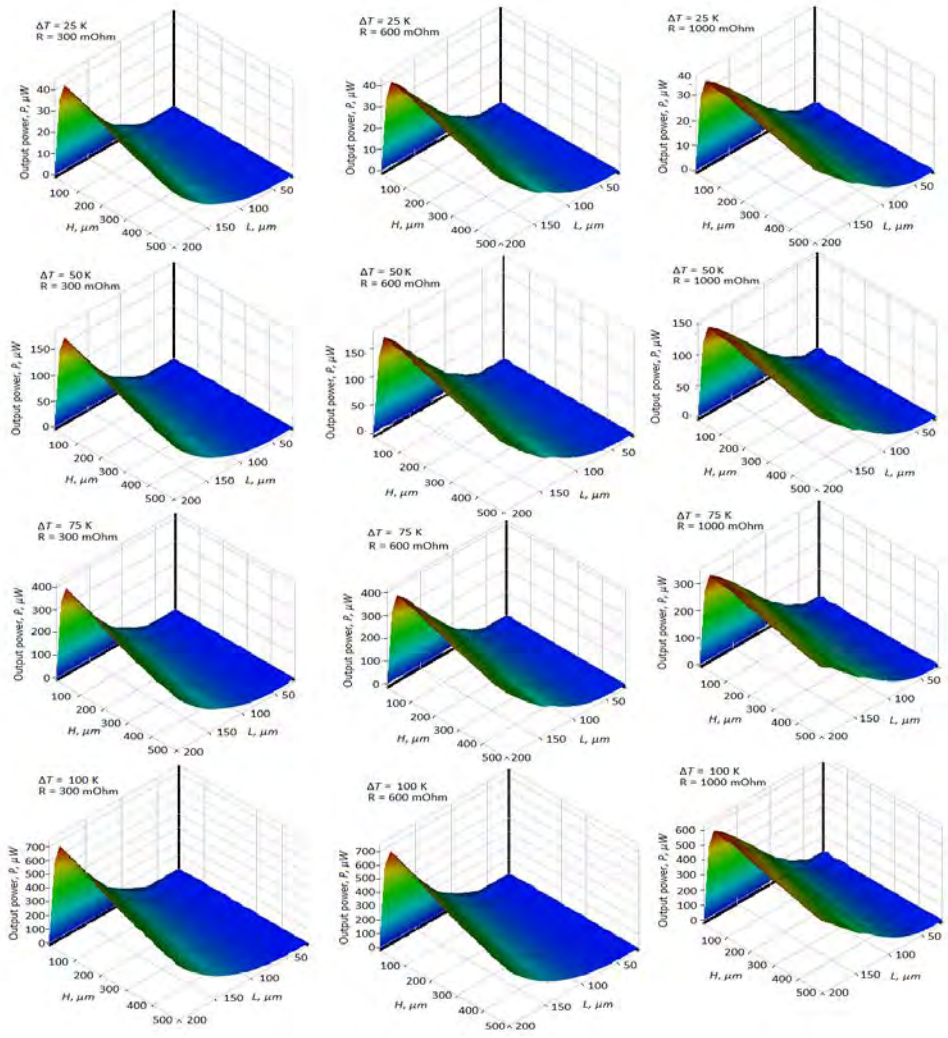


Fig. 2 μTEG output power dependences on thermal elements linear dimensions with different external load values and temperature ranges

[8] G. Zeng, J-H. Bahk, J.E. Bowers, J.M.O. Zide, A.C. Gossard, Z. Bian, R. Singh, A. Shakouri, W. Kim, S.L. Singer, and A. Majumdar "ErAs: $(InGaAs)_{1-x}(InAlAs)_x$ alloy power generator modules," Appl. Phys. Lett, 91, 263510 (2007).

[9] L. Tzounisa, M. Liebscher, R. Fuge, A. Leonhardt, V. Mechtcherine "P- and n-type thermoelectric cement composites with CVD grown p- and n-doped carbon nanotubes: Demonstration of a structural thermoelectric generator." Energy and Buildings, vol. 191, 15 May 2019, pp. 151-163.

[10] M.S. Yenuchenko, A.S. Korotkov, D.V. Morozov and M.M. Pilipko, "A Switching Sequence for Unary Digital-to-Analog Converters Based on a Knight's Tour," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 6, pp. 2230-2239, June 2019.

[11] D.B. Akhmetov, A.S. Korotkov, D.V. Morozov, M.M. Pilipko, I. A. Rumyancev, "Radio Frequency Identification System of Internet of Things Based on CMOS Integrated Circuits," Proc. 15th EWDTS, 29.09-02.10.2017, Novi Sad, Serbiay, pp.603-605.

[12] D.B. Akhmetov, A.S. Korotkov, I.A. Rumyancev. 2.4-2.5 GHz Fractional-N Frequency Synthesizer with Integrated VCO in 0.18 um CMOS for RFID Systems // Proc. IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), 22-23 October 2018, St. Petersburg, Russia, pp.64-68.

[13] A.F. Ioffe Semiconductor Thermoelements and Thermoelectric Cooling, Infosearch, London, 1957.

[14] V.V. Loboda, A.S. Korotkov, S.V.Dzyubanenko, E.M.Bakulin, "Design of the Thin-Film Thermoelectric Generator for Low-Power Applications," Russian Microelectronics, 2019, vol. 48, no. 5, pp. 326–334.

[15] R. Buslaev, V. Loboda, "Simulation of Uni-Leg thermoelectric generator," Proc. IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), 22-23 October 2018, St. Petersburg, Russia, pp. 27-31.

[16] A. Korotkov, V. Loboda, A. Feldhoff, D. Groeneveld, "Simulation of Thermoelectric Generators and Its Results Experimental Verification," Proc. IEEE International Symposium on Signals, Circuits and Systems (ISSCS 2017), 13-14 July 2017, Iasi, Romania, 2017.

[17] A.S. Korotkov, V.V. Loboda, "Simulation and Design of Thin-Film Thermoelectric Generators," Proc. International Symposium on Fundamentals of Electrical Engineering (ISFEE-2018), 1-3 November 2018, Bucharest, Romania, 2018.

[18] R. Buslaev, A. Galitskaya, V. Loboda, "Simulation of Flexible Thermoelectric Generators Based on Bi2Te3/Sb2Te3 Synthesized by Electrochemical Deposition Method," Proc. IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), 17-18 October 2019, St. Petersburg, Russia, pp. 54-57.

[19] G.J. Snyder, E.S. Toberer "Complex thermoelectric materials." Nature materials, Vol. 7, February 2008, pp. 105–114.

[20] A. Piggott "Detailed Transient Multiphysics Model for Fast and Accurate Design, Simulation and Optimization of a Thermoelectric Generator (TEG) or Thermal Energy Harvesting Device," Journal of Electronic Materials, 2019, Vol. 48, pp. 5442–5452.

[21] S. Ferreira-Teixeira, A.M. Pereira Geometrical "Optimization of a Thermoelectric Device: Numerical Simulations," Energy Conversion and Management, Vol. 169, 2018, pp. 217-227.

[22] J. Dongxu, W. Zhongbao, J. Pou, S. Mazzoni, S. Rajoo, F. Romagnoli "Geometry Optimization of Thermoelectric Modules: Simulation and Experimental Study," Energy Conversion and Management, Vol. 195, 2019, pp. 236-243.

Hardware Obfuscation Techniques on FPGA-Based Systems

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

Olha Ponomarenko
Department of Electronic Computers
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
olha.ponomarenko@nure.ua

Oksana Kotkova
Department of Electronic Computers
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
oksana.kotkova@nure.ua

***Abstract*—There is a great variety of hardware Trojan detection and prevention approaches. However, state of art approaches cannot provide a full guarantee that an integrated circuit or complex electronic system is free of hardware Trojan. We introduced a reference monitor obfuscation approach on the base of formal transformations of structural system models. The approach ensures development of secure systems, operating in the presence of hardware Trojans. The reference monitor obfuscation ensures the main key reference monitor properties: must be non-bypassable and tamper-proof. This concept can be used as a prevention countermeasure against hardware Trojans at the following steps of development cycle of integrated circuit on the FPGA platform: prevention at design, prevention at fabrication, and prevention at post-fabrication. The paper demonstrates an implementation of reference monitor obfuscation approach by physical modeling on FPGA-based systems.**

Keywords—hardware security, structural model, trusted computing base, hardware design obfuscation, FPGA platform

I. INTRODUCTION

Sometimes, a secure system is considered as an information processing system that includes a particular set of security features. From our viewpoint the availability of security mechanism is only a necessary condition and cannot be considered as a criterion for the system security from real threats. A more accurate definition of a secure system describes a secure system as a system that ensures the security of the processed information and maintains its operability under the conditions of exposure to it of a certain number of threats [1]. Requirements for such systems include access control mechanism as a mandatory component; ability to formally assess system security.

Computer architectures are layered. There are two concepts of security mechanism implementation: security system is placed at any layer; layered kernel-based security architecture.

Security system at any architectural layer. In this case, designers make decision about layer for security mechanism. At the same time, they should consider protection of layers below. Embedding security mechanisms at various levels has several serious disadvantages.

- Security mechanism of chosen layer relies on reliable security mechanism of the layers below. For example. Operating system access controls typically rely on reliable hardware. Thus, it is possible to attack from the lower layer to the upper layer.
- Rising from hardware through the operating system and middleware to application layer, the processes of interaction of all layers of a system become more complex and secure mechanisms less reliable.

Layered Kernel-based security architecture. It is natural to assume that all security mechanisms should themselves be protected from any attacks. Otherwise, it will be difficult to talk about the reliability of protection. Therefore, it is logical to combine all security mechanisms into so-called Trusted Computing Base (TCB). A reliable computing base is an abstract concept denoting a set of protective mechanisms of a computing system, including software and hardware components, responsible for maintaining a security policy [2], [3]. The [4] outlines some of the basic features of TCB such as: maintaining the confidentiality, integrity and accessibility of data on a system, enforcing the system's security policy, and protection against any forms of system infiltration.

Designers of hardware and hardware-software secure systems use TrustZone technology [5]. This technology provides a security framework possessing features of TCB and enables a system to counter many of the specific threats that it will experience.

A TCB consists of one or more components that together are responsible for implementing a unified security policy within the system. The ability of the TCB to correctly implement a unified security policy depends primarily on the mechanisms of the TCB itself, as well as on the correct management by the system administration. Thus, TCB performs a dual task: it supports the implementation of security policies and is the guarantor of the integrity of protection mechanisms. The structure of TCB includes such components: reference monitor (RM) and security kernel (SK).

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

A RM is an access control concept of an abstract machine that mediates all accesses to objects by subjects [6], [7]. A TCB is RM plus other security mechanisms. Designers using the reference monitor have the ability to include security aspect into design process. The RM must be non-bypassable, evaluable, always invoked, and tamper-proof. SK is software and hardware implementation of RM. Designers using the reference monitor have the ability to include security aspect into design process.

In [8] authors consider in detail issues of building a multi-level (multi-layer) OS security architecture. Designing a secure OS is carried out using the concepts of TCB, Reference Monitor and Security Kernel.

It should be noted that in [9] several disadvantages are considered related to the practical use of the concept of TCB. So, there is no answer to the question of how to include components that are responsible for the implementation of a unified security policy within the system and how to determine the boundaries of the TCB.

Approaches which can be used as countermeasures against HTs are divided into two groups: detection and prevention. Detection and prevention approaches are classified in [10]. Limitations of countermeasures are following. All countermeasures against HTs are designed to defend against only a subset of the possible HT attacks that they may experience. Defending against all possible attacks is an impossible task. The best that can be achieved is design and fabrication of secure systems, operating in the presence of HTs.

A promising set of approaches against HTs is hardware obfuscation-based approaches [11]. A detailed classification of obfuscation approaches is given in [12].

In the paper [13], the authors consider a hardware obfuscation approaches for two levels of IC representation: layout-level and netlist-level. Logic obfuscation is demonstrated in [14]. A promising approach to obfuscate complex electronic system is addressed in [15]. This approach has a disadvantage: implementation of obfuscation requires considerable hardware overhead.

The main goal of the proposed research is to develop design technique that can effectively resist or mitigate security threats at untrusted stages of the IC life-cycle. The main principle of the proposed technique is design obfuscation method on FPGA-based systems.

II. Ease Reference Monitor Obfuscation Techniques on FPGA-Based Systems

This section is devoted to the Reference Monitor Obfuscation on the base of system model aggregation. A system S contains the elements C_i, where i = 0,1,…,6. Let's consider the formal model of a complex system structure [16].

Consider a system, which structure is represented in Fig. 1. Each system component C_j is represented by sets of input $\left[X_i^{(j)}\right]_1^m$ and output $\left[Y_l^{(j)}\right]_1^r$ contacts, where $i = \overline{0,m_j}$, $l = \overline{0,r_j}$. To simplify, we will denote $m = m_j$, $r = r_j$. A single-valued operator [17]

$$Y_l^{(k)} = R\left(X_i^{(j)}\right) \quad (1)$$

is proposed to describe a formal structure model of system. Where the domain of the operator is the set $\bigcup_{j=0}^{N}\left[X_i^{(j)}\right]_1^m$ and the codomain of the operator is the set $\bigcup_{k=0}^{N}\left[Y_l^{(k)}\right]_1^r$.

The operator (1) is represented in a tabular form. The Table 1 depicts the operator values for the system under consideration

TABLE I. The Operator R of the Elements Connections for the System S

j \ i	1	2	3	4	5
0	1,1	3,1	4,1	5,1	6,2
1	0,1				
2	1,3	0,2	0,3		
3	1,2	2,1			
4	3,2	2,1	2,2		
5	2,2				
6	5,2	0,4			

A row of the Table 1 corresponds to an element of system S. A column corresponds to an input contact of element. The intersection of row j and column i gives a pair (k, l). Where k is the number of element and l is its output contact to which the input contact i of element j is connected.

Description of the approach proposed. System S consists of $S_{\mu 0} = \{C_0, C_3, C_4, C_5, C_6\}$ and $S_{\mu 1} = \{C_1, C_2\}$. The subsystems are indicated by a dashed line in Fig. 1. Let the subsystem $S_{\mu 1}$ plays a role of RM. Three steps of proposed approach: determination of additional fictitious contacts, construction of the operators of elements connections R for $S_{\mu 0}$ and $S_{\mu 1}$ and RM obfuscation procedure.

Step 1. Determination of additional fictitious contacts. The major idea of this stage is as follows. Each of the subsystems ($S_{\mu 0}$ or $S_{\mu 1}$) can be an element: the subsystems $S_{\mu 0}$ is an element of a system $S^1 = \{C_1, C_2, S_{\mu 0}\}$, the subsystems $S_{\mu 1}$ is an element of a system $S^2 = \{C_0, C_3, C_4, C_5, C_6, S_{\mu 1}\}$.

An access of elements of subsystem $S_{\mu 1}$ to elements of subsystem $S_{\mu 0}$ is possible only through its fictitious input $X^{(0)\mu 1}$ and output contacts $Y^{(0)\mu 1}$ (Fig. 1). The contacts $X^{(0)\mu 1}$ are linked to output contacts of elements of $S_{\mu 1}$. The contacts $Y^{(0)\mu 1}$ are linked to input contacts of elements of $S_{\mu 1}$.

An access of elements of subsystem $S_{\mu 0}$ to elements of subsystem $S_{\mu 1}$ is possible only through its fictitious input $X^{(\mu 1)}$ and output contacts $Y^{(\mu 1)}$ (Fig. 1). The contacts $X^{(\mu 1)}$ are linked to output contacts of elements of $S_{\mu 0}$. The contacts $Y^{(\mu 1)}$ are linked to input contacts of elements of $S_{\mu 0}$.

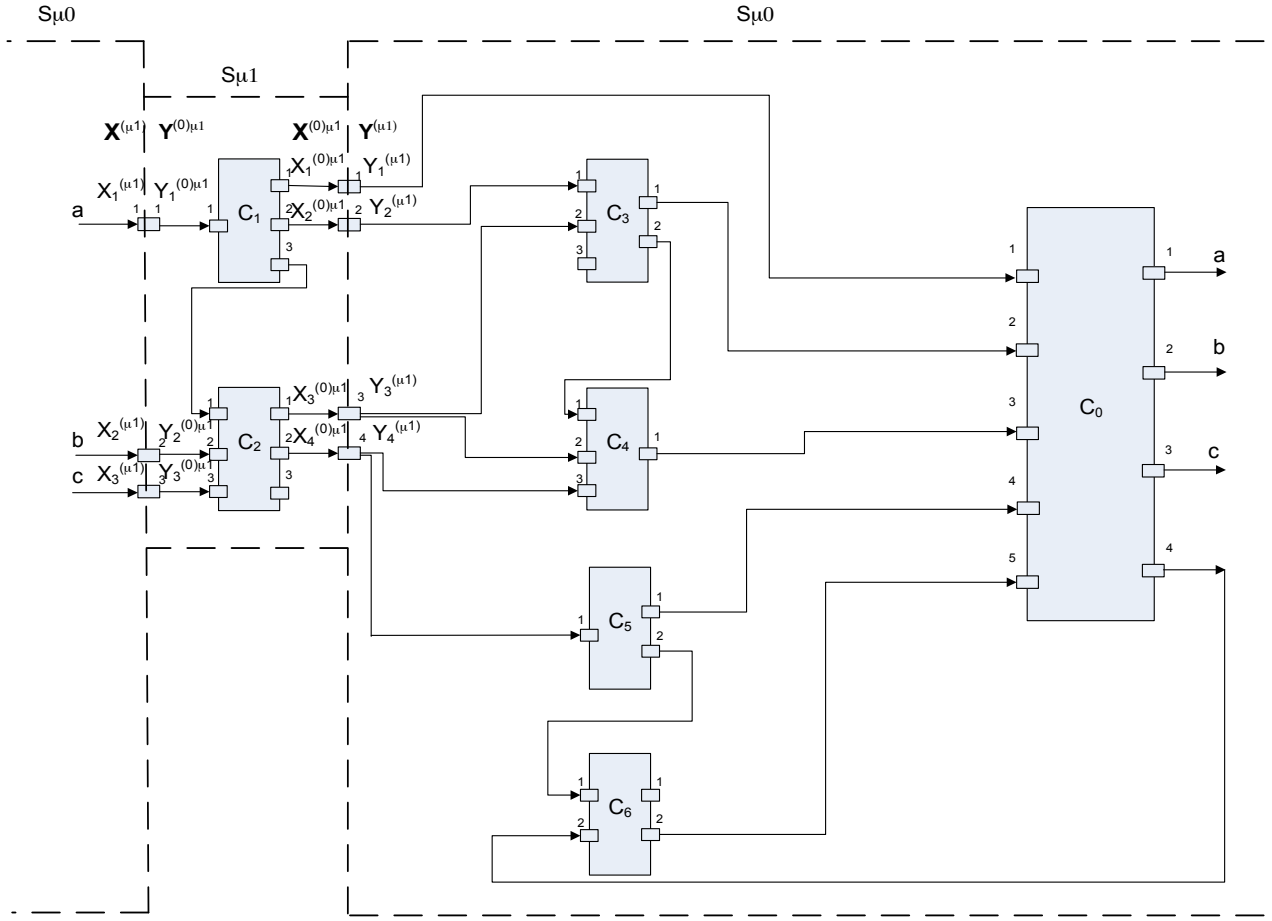


Fig. 1. The structure of system S.

Let consider a procedure of fictitious contacts construction for the system under consideration. As declared above, a formal description of element C_j is sets: $\left[X_i^{(j)}\right]_1^m$ and $\left[Y_l^{(j)}\right]_1^r$, where $i=\overline{0,m_j}$, $l=\overline{0,r_j}$. For each $Y_l^{(j)}$, the operators below are used to construct fictitious contacts:

$$\begin{aligned} Y_l^{(\mu)} &= Q_\mu(Y_l^{(j)}), \\ X_i^{(0)\mu} &= Q'_\mu(Y_l^{(j)}) \end{aligned} \tag{2}$$

The operators Q_μ and Q'_μ are represented by a fictitious contact numbering table. The numbering table of the fictitious contacts of $S_{\mu 1}$ has a form of Table 2.

TABLE II. THE NUMBERING TABLE OF OPERATORS Q_μ AND Q'_μ OF THE SUBSYSTEM $S_{\mu 1}$

$Y_l^{(j)}$	(j,l)	1,1	1,2	2,1	2,2
$Q_{\mu 1}$	$\mathbf{Y}^{(\mu 1)}$	$Y_1^{(\mu 1)}$	$Y_2^{(\mu 1)}$	$Y_3^{(\mu 1))}$	$Y_4^{(\mu 1)}$
$Q_{\mu 1}'$	$\mathbf{X}^{(0)\mu 1}$	$X_1^{(0)\mu 1}$	$X_2^{(0)\mu 1}$	$X_3^{(0)\mu 1}$	$X_4^{(0)\mu 1}$

For each element $X_i^{(j)}$, the operators below are used to construct fictitious contacts:

$$\begin{aligned} X_i^{(\mu)} &= P_\mu(X_i^{(j)}), \\ Y_l^{(0)\mu} &= P'_\mu(X_i^{(j)}) \end{aligned} \tag{3}$$

The operators P_μ and P'_μ are represented by a fictitious contact numbering table. The numbering table for P_μ and P'_μ of $S_{\mu 1}$ has a form of Table 3.

TABLE III. THE NUMBERING TABLE OF OPERATORS P_μ AND P'_μ OF THE SUBSYSTEM $S_{\mu 1}$

$X_i^{(j)}$	(j, i)	1,1	2,2	2,3
P_μ	$i_\mu(X_i^{(\mu 1)})$	1	2	3
P'_μ	$l_{0\mu}(Y_l^{(0)\mu 1})$	1	2	3

The operators (2) and (3) are some procedures that enable numbering fictitious contacts for the subsystem $S_{\mu 1}$. Similar to the subsystem $S_{\mu 1}$ the operators (2) and (3) are implemented to the subsystem $S_{\mu 0}$ that enable numbering of fictitious contacts for the subsystem $S_{\mu 0}$. The result is demonstrated in Fig. 1.

Step 2. The operators of elements connections R for $S_{\mu 0}$ and $S_{\mu 1}$. Denote the operator R for a case of subsystem S_μ by R_μ. The major idea of this step is following. An operator R_μ consider a subsystem S_μ as an element with fictitious contacts just like other elements of system S under consideration. The procedure of constructing the operator R_μ is defined by the expression (4).

The operator (4), like the operator (1), assigns the output contact $Y_l^{(k)}$ to the input contact $X_i^{(j)}$, taking into consideration, that the domain of R_μ consists of three subsets. The operator (3) is represented in a tabular form.

$$Y_l^{(k)} = \begin{cases} R\left(X_i^{(j)}\right) \text{for } X_i^{(j)} \in \bigcup\limits_{C_j \in S_\mu} \bigcup\limits_{C_k \in S_\mu} \left[X^{(j,k)}\right] \\ P_\mu^1\left(X_i^{(j)}\right) \text{for } X_i^{(j)} \in \bigcup\limits_{C_j \in S_\mu} \bigcup\limits_{C_k \notin S_\mu} \left[X^{(j,k)}\right] \\ (Q_\mu^1)^{-1}\left(X_i^{(0)\mu}\right) \text{for } X_i^{(0)\mu} \in \left[X_i^{(0)\mu}\right]_l^m \end{cases} \quad (4)$$

Operator values $R_{\mu 1}$ for the system $S_1 = \{C_1, C_2, S_{\mu 0}\}$ are presented in Table 4.

TABLE IV. THE OPERATOR $R_{\mu 1}$ FOR THE SUBSYSTEM $S_{\mu 1}$

i / j	1	2	3	4
0	1,1	1,2	2,1	2,2
1	$0,Y_1^{(0)\mu 1}$			
2	1,3	$0,Y_2^{(0)\mu 1}$	$0,Y_3^{(0)\mu 1}$	

The row of the table header corresponds to the number of fictitious input contacts of $S_{\mu 0}$ and number of inputs of C_1 and C_2. The subsystem $S_{\mu 0}$ is represented by the row 0. The elements C_1 and C_2 are represented by rows 1 and 2, respectively. The maximum number of inputs among C_1, C_2 and $S_{\mu 0}$ determines the number of columns. In the table, $Y_l^{(0)\mu}$ are output contacts of $S_{\mu 0}$. At the intersection of rows and columns there are pairs of numbers (k, l). Where k is the number of element, l is the number of its output contact.

Values of operator $R_{\mu 0}$ for $S^2 = \{C_0, C_3, C_4, C_5, C_6, S_{\mu 1}\}$ are given in Table 5.

TABLE V. THE OPERATOR $R_{\mu 0}$ FOR THE SUBSYSTEM $S_{\mu 0}$

i / j	$1(X_1^{(\mu 1)})$	$2(X_2^{(\mu 1)})$	$3(X_3^{(\mu 1)})$	4	5
0	$S_{\mu 1}$,1	3,1	4,1	5,1	6,2
3	$S_{\mu 1}$,2	$1_{\mu 1}$,3			
4	3,2	$S_{\mu 1}$,3	$S_{\mu 1}$,4		
5	$S_{\mu 1}$,4				
6	5,2	0,4			
$S_{\mu 1}$	0,1	0,2	0,3		

The row of the table header contains numbers of fictitious input contacts of $S_{\mu 1}$ and numbers of inputs of C_0, C_3, C_4, C_5, and C_6. The rows 0,3,4,5 and 6 correspond to the elements C_0, C_3, C_4, C_5, and C_6, respectively. The subsystem $S_{\mu 1}$ is represented by the row $S_{\mu 1}$. The maximum number of inputs among C_0, C_3, C_4, C_5, C_6 and $S_{\mu 1}$ determines the number of columns. At the intersection of rows and columns there are pairs of numbers (k, l). Where k is the number of element, l is the number of its output contact.

Step 3. Let consider RM obfuscation procedure proposed in the work. The concept of RM divides a design into different subsystems $S_{\mu 0}$ and $S_{\mu 1}$. The untrusted subsystem $S_{\mu 0}$ (a design) does not have access to the trusted subsystem $S_{\mu 1}$ (RM). These subsystems can be designing and manufacturing by different teams. That enables obfuscating RM and minimizing risks of Trojans attacks.

III. PHYSICAL MODELING

To demonstrate a feasibility of the method described above, we conducted physical modeling its steps targeted to RM obfuscation. For practical test of design method we used tool set Xilinx Spartan-3E Starter Kit. The Spartan 3E Starter evaluation board was used as a platform for design. To write a VHDL code the free Xilinx Integrated Software Environment (ISE) 14.7 was used. To demonstrate a major concept of the method structural description style was used. Several motivations make structural description usage more preferable: structural description based on system structural model is represented in terms of the interconnection of its subsystems instead of focusing on components functionality; structural description effectively solves the problem of RM obfuscation using a structure of design consisting of two subsystems and fictitious contacts.

An implementation of the design method in the VHDL environment. The structural VHDL model describes two-level decomposition of the design into subsystems $S_{\mu 0}$ and $S_{\mu 1}$. Each subsystem is a structural level VHDL module (entity). The module includes a declaration of elements C_i and their behavioral description. The modules interact by means of fictitious contacts, which are connected to ports of modules by corresponding signals. As noted above, the subsystem $S_{\mu 0}$ implements the main functions of the project, and the subsystem $S_{\mu 1}$, consisting of the elements C_1 and C_2 is a RM, that controls accesses of elements of $S_{\mu 0}$ to each other. In our case, the element C_1 controls the accesses of C_0 to the element C_3, and the element C_2 controls the accesses of C_0 to the element C_4. Element C_0 from the output $Y_l^{(0)}$ sends a request to element C_3 for reading information. In the initial state, at the output $Y_l^{(3)}$, state Z (state of high impedance) is set. Element C_1 determines the right of access of element C_0 to element C_3 and grants or does not grants an access. If there is a right of access, element C_0, at the input $X_1^{(0)}$, reads information from output $Y_l^{(3)}$ of element C_3. The access control of element C_0 to element C_4 is carried out be the element C_2 in a similar manner.

It is evident, the subsystems can be designing and manufacturing separately with their subsequent integration within the framework of a single design.

IV. Conclusion And Future Work

We introduced the reference monitor obfuscation approach on the base of formal transformations of structural models The RM obfuscation technique ensures the key property of access control mechanism: it must be non-bypassable and tamper-proof. This technique can be used as a prevention countermeasure against HTs at the following steps of a modern FPGA-based IC life cycle such as: prevention at design, prevention at fabrication, and prevention at post-fabrication. RM obfuscation concept allows to develop secure systems, operating in the presence of HTs.

Future investigations are necessary to evaluate the resilience of approaches based on RM obfuscation. A promising direction of future studies is design technologies, based on theoretically proofed guarantees of system security.

References

[1] M. Bishop, Computer Security: art and science, Addison Wesley, ISBN 0-201-44099-7, 2002.

[2] Department of Defense trusted computer system evaluation criteria, Dept. of Defense, 1985.

[3] M. Heckman and R. Schell, "Using Proven Reference Monitor Patterns for Security Evaluation. Information," 7, 23, 10.3390/info7020023, 2016.

[4] B. Lampson, M. Abadi, M. Burrows and E. Wobber, "Authentication in Distributed Systems: Theory and Practice," ACM Transactions on Computer Systems, 1992, p. 6.

[5] ARM Security Technology. Building a Secure System using TrustZone® Technology 2005-2009 ARM Limited, PRD29-GENC-009492C.

[6] J. Anderson, "Computer Security Technology Planning Study," Technical Report ESD-TR-73–51, Electronic Systems Division, Hanscom Air Force Base, Hanscom, MA, 1974.

[7] C. Irvine, "The Reference Monitor Concept as a Unifying Principle in Computer Security Education," in Proceedings of the first world conference on information security education, 1999, pp. 27–37.

[8] C. Pfleeger, S. Pfleeger and J. Margulies, Security in Computing, 5th ed., Prentice Hall, pp. 1043, 2015.

[9] B. Blakley and D. M. Kienzle, "Some Weaknesses of the TCB Model," in Proceedings of the 1997 IEEE Symposium on Security and Privacy, IEEE Computer Society Press, 1997.

[10] J. Francq, "Hardware Trojans Detection Methods," Cassidian Cybersecurity - All rights reserved, in TRUDEVICE, 2013, pp. 36–40.

[11] A. Sengupta, D. Roy, S. Mohanty and P. Corcoran, "DSP design protection in CE through algorithmic transformation based structural obfuscation," IEEE Transactions on Consumer Electronics, vol. 63, no. 4, 2017, pp. 467–476.

[12] D. Forte, S. Bhunia and M. M. Tehranipoor, Hardware Protection through Obfuscation, Springer International Publishing, 2017 (Chapter 2, Fareena Saqib and Jim Plusquellic, VLSI Test and Hardware Security Background for Hardware Obfuscation).

[13] G. T. Becker, M. Fyrbiak and C. Kison, "Hardware obfuscation: Techniques and open challenges," Book Chapter, Foundations of Hardware IP Protection, Springer International Publishing, 2017, pp. 105–123.

[14] H. Li, Q. Liu and J. Zhang, "A survey of hardware Trojan threat and defense," Integration, the VLSI Journal, vol. 55, 2016.

[15] C. Pilato, F. Regazzoni, R. Karri and S. Garg, "TAO: Techniques for Algorithm-Level Obfuscation during High-Level Synthesis," in Proceedings of the the 55th Annual Design Automation Conference, June 2018.

[16] R. Haggaty, Discrete Mathematics for Computing, Edinburgh: Pearson Education Limited, 2002.

[17] V. Gorbachov, D. Sytnikov, O. Ryabov, A. K. Batiaa and O. Ponomarenko, "Dimension Reduction for Network Systems Using Structure Model Aggregation," International Journal of Design & Nature and Ecodynamics, vol. 15, no. 1, February 2020, pp. 13–23.

Weighted Total Least Squares for Frequency Estimation of Real Sinusoids Based on Augmented System

Dmitriy V Ivanov
dept. Mechatronics
Samara State University of Transport
Samara, Russia
dvi85@list.ru

Alexander I. Zhdanov
dept. Higher Mathematics
Samara State Technical University
Samara, Russia
ZhdanovAleksan@yandex.ru

***Abstract*— Estimating the frequencies of sinusoids based on the weighted total least squares method allows for obtaining more accurate estimates compared to the ordinary least squares method. The most common algorithm for solving the total least squares problem is the algorithm based on the singular value decomposition (SVD) of a matrix. This algorithm has a high computational complexity. An alternative method for solving the total least squares problem is the biased normal systems approach. The paper proposes augmented system of linear algebraic equations equivalent to a weighted normal biased system of equations. and an estimate of the frequencies of sinusoids from discrete measurements with based on the proposed augmented system. The simulation results show that the frequency estimates based on the augmented systems are comparable in accuracy with the solution of the total least squares problem based on the singular value decomposition of the matrix.**

Keywords—Total least square, resolution, frequency estimation, ill condition, real sinusoids.

I. INTRODUCTION

The problem considered in the paper is one of the problems of spectral analysis [1]. In many applications, especially in the field of communications, radar, sonar, geophysics, seismology, the signals under consideration can be well described by the sum of sinusoids with noise [2]. A large number of papers [1] - [5] are devoted to the problem of estimating the frequencies of sinusoids from a finite number of discrete noisy measurements because of its wide application in science and technology.

Although there are a large number of frequency estimation methods, they can be classified as nonparametric or parametric approaches. Nonparametric frequency methods are based on the application of the Fourier transform. Nonparametric methods do not make any assumptions about the pattern of the observed sequence data. The resolution or ability to resolve closely spaced frequencies using nonparametric methods is fundamentally limited by the length of the available data. Alternatively, a parametric approach is used, which assumes that the signal satisfies a generative model with a known functional form. The parametric approach allows for higher resolution. The following parametric estimates of frequency are known: maximum likelihood (MP) [6], nonlinear least squares (NLS) [7], total least squares (TLS) [8, 9], instrumental variables (IV) [10], Yule-Walker equations [11], iterative filtering [12] and subspace methods such as truncated singular value decomposition, MUSIC and ESPRIT [4].

In fact, with additive white Gaussian noise, the MP and NLS methods are equivalent, and both are statistically efficient, but their computational complexity is very high. On the other hand, the rest of the above-mentioned parametric methods use linear prediction (LP) of sinusoidal signals and provide a suboptimal efficiency estimate, but their computational complexity is less.

The total least squares is used in many methods for estimating the frequencies of sinusoids. In [8], an estimate for sinusoids is proposed based on total least squares using use linear prediction (LP) of sinusoidal signals. Weighted versions of the algorithm [8] are discussed in [13, 14]. Also, the method of total least squares is used for methods based on the higher order Yule-Walker equations [15, 16]. These algorithms do not take into account the symmetry of the coefficients of the estimated coefficients LP of real sinusoidal signals.

In [17], a weighted algorithm is proposed for estimating coefficients taking into account the symmetry of the coefficients of the LP model of real sinusoidal signals. The proposed algorithm requires compute the generalized eigenvector at each iteration. This is a complex nonlinear computational problem [18].

The fundamental limitation for the resolution of parametric methods is the computational stability of the algorithms and the unbiasedness of the estimates obtained. This is because at close frequencies of the estimated sinusoids the problem becomes ill-conditioned. It is known that the use of total least squares, allows you to increase the resolution in comparison with the ordinary least squares.

The most common algorithm for solving the total least squares problem is the algorithm based on the singular value decomposition of a matrix. This algorithm has a high computational complexity. An alternative method for solving the total least squares problem is the biased normal systems approach. In [19], it was noted that the solution of the least

squares problem in the form of a biased normal system may be preferable in some cases.

One of the effective methods for solving a biased normal system is to convert to an augmented equivalent system [20]. An augmented system equivalent to a weighted biased system is obtained in the paper. On the basis of the proposed augmented system, an algorithm for estimating the frequencies of sinusoids is implemented.

II. Problem Statement

The problem of estimating the frequencies of sinusoids is formulated as follows. Discrete measurements with noise are given

$$y(n) = x(n) + e(n),\ n = 0,1,\ldots,N-1 \tag{1}$$

where

$$x(n) = \sum_{l=1}^{L} A_l \sin(\omega_l n + \varphi_l),\ l = 1,2,\ldots,\mathrm{L} \tag{2}$$

A_l, $\omega_l = 2\pi f_l \in (0,\pi)$ and $\varphi_l \in [0,2\pi)$ are unknown amplitude, frequency, and phase l-th sinusoid. $e(n)$ is an additive white noise with zero mean unknown variance σ_e^2. The number L is known a priori.

It is required to estimate the frequencies $\hat{\omega}_l$ from the noisy sequences of observations $\{y(n)\}$.

III. Linear Prediction Property

Estimating frequencies ω_l from a noisy signal is a difficult task because the frequencies ω_l are included in (2) non-linearly.

The signal $x(n)$ can be uniquely represented as a linear combination $2\mathrm{L}$ of previous values

$$x(n) = -\sum_{i=1}^{2L} a_i x(n-i), \tag{3}$$

where a_l are constant coefficients. Frequencies $\omega_l \in (0,\pi)$ are related to coefficients a_l by the following equation [9]:

$$\sum_{i=0}^{2L} a_i \exp(-j\omega_l i) = 0, \tag{4}$$

$a_0 = 1$, $a_i = a_{2L-i}$, $j = \sqrt{-1}$.

From solving the equation

$$\sum_{i=0}^{2L} a_i z^i = 0, \tag{5}$$

frequencies ω_l can be estimated as the phase $\exp(\pm j\omega_l)$ of the roots of equation (5). Thus, the problem of estimating frequencies can be reduced to estimating the coefficients a_i included in the equation linearly.

Using the symmetry property of the coefficients, (3) is represented as follows

$$\begin{gathered} -x(n) - x(n-2L) = \\ = \sum_{i=1}^{L-1} a_i \left(x(n-i) - x(n-2L+i) \right) + a_L x(n-L), \\ y(n) = x(n) + e(n), \end{gathered} \tag{6}$$

Let's define an expression for the prediction error

$$\begin{gathered} \xi(n) = e(n) + e(n-2L) + \\ + \sum_{i=1}^{L-1} a_i \left(e(n-i) - e(n-2L+i) \right) + a_L e(n-L). \end{gathered} \tag{7}$$

Let us write (6) in matrix form

$$\mathbf{Y} = \mathbf{X}\mathbf{a}, \tag{8}$$

where $\mathbf{Y} = \left(-y(N-1) - y(N-2L-1), \ldots, -y(2L) - y(0) \right)^T$, $\mathbf{a} = (a_1,\ldots,a_l)^T$,

$$\mathbf{X} = \begin{pmatrix} y(N-2) + y(N-2L) & \ldots & y(N-L-1) \\ y(N-3) + y(N-2L-1) & \ldots & y(N-L-2) \\ \vdots & \ddots & \vdots \\ y(2L-1) + y(1) & \ldots & y(L) \end{pmatrix}^T.$$

IV. Weighted Total Least Squares For Frequency Estimation

The frequency estimation based on total least squares is unbiased, but not effective. Modifications of weighting algorithms are used to increase the accuracy of estimates. Consider the solution to the weighted total least squares problem:

$$\mathbf{W}\mathbf{Y} = \mathbf{W}\mathbf{X}\mathbf{a}, \tag{9}$$

where $\mathbf{W}$ is positive definite matrix.

In case the noise $e(n)$ is Gaussian distribution. Weighting matrix use selection

$$\mathbf{W} = \sigma_e^2 \mathrm{E}\left(\xi\xi^T \right)^{-1}, \tag{11}$$

allows you to get the maximum likelihood estimates.

In [17], a frequency estimation is proposed based on finding a generalized eigenvector. We get the frequency estimation by the weighted total least squares method based on the solution of the augmented system. Solving weighted total least square is described by minimizing the generalized Rayleigh ratio [21]

$$\min_{\mathbf{a}} \frac{\left(\mathbf{Y}-\mathbf{X}\mathbf{a}\right)^T \mathbf{W}\left(\mathbf{Y}-\mathbf{X}\mathbf{a}\right)}{1+\mathbf{a}^T\mathbf{G}\mathbf{a}}. \quad (12)$$

where $\mathbf{G}=\dfrac{\mathrm{E}\left(\tilde{\mathbf{E}}^T\mathbf{W}\tilde{\mathbf{E}}\right)}{2\sigma_e^2}$,

$$\tilde{\mathbf{E}}=\begin{pmatrix} e(N-2)+e(N-2L) & \dots & e(N-L-1) \\ e(N-3)-e(N-2L-1) & \dots & e(N-L-2) \\ \vdots & \ddots & \vdots \\ e(2L-1)-e(1) & \dots & e(L) \end{pmatrix}^T.$$

In [17], expressions for matrices $\mathbf{W}^{-1}$ и $\mathbf{G}$ are also obtained in explicit form. The matrix $\mathbf{W}^{-1}$ is a banded Toeplitz matrix

$$\mathbf{W}^{-1}=\begin{pmatrix} W_0^{-1} & W_1^{-1} & \dots & W_{2L}^{-1} & 0 & \dots & 0 \\ W_1^{-1} & W_0^{-1} & W_1^{-1} & \dots & W_{2L}^{-1} & \dots & 0 \\ \vdots & & \vdots & & & \vdots & \vdots \\ 0 & \dots & 0 & \dots & \dots & W_1^{-1} & W_0^{-1} \end{pmatrix},$$

where $W_0^{-1}=\mathrm{E}\left(\tilde{\xi}(n)\tilde{\xi}(n)\right)=\left(\bar{a}_L^2+2\sum_{i=1}^{L-1}\bar{a}_i^2\right)$,

$$W_1^{-1}=\mathrm{E}\left(\tilde{\xi}(n)\tilde{\xi}(n+1)\right)=2\sum_{i=1}^{L-1}\bar{a}_i\,\bar{a}_{i+1},\dots,$$

$$W_{2L}^{-1}=\mathrm{E}\left(\tilde{\xi}(n)\tilde{\xi}(n+2L)\right)=1,$$

$$\mathbf{G}=\begin{pmatrix} G_0+G_{2L} & G_1+G_{2L-1} & \dots & G_L+G_{2L-1} & G_L \\ G_1+G_{2L-1} & G_0+G_{2L-2} & \dots & G_{L-1}+G_{2L-3} & G_{L-1} \\ G_2+G_{2L-2} & G_1+G_{2L-3} & \dots & G_{L-2}+G_{L-3} & G_{L-2} \\ \vdots & \vdots & & \vdots & \vdots \\ G_L & G_{L-1} & \dots & G_1 & G_0/2 \end{pmatrix},$$

$$G_j=\sum_{i=1}^{N-2L-j} W_{i,i+j}.$$

Remark. For banded Toeplitz matrices, there are fast algorithms for finding inverse matrices, which can significantly reduce the complexity of the [22, 23] algorithm.

We introduce a new vector of variables

$$\tilde{\mathbf{a}}=\mathbf{H}_G\mathbf{a}, \quad (13)$$

where $\mathbf{G}=\mathbf{H}_G^T\mathbf{H}$.

Let us write criterion (12) using the new vector of variables

$$\min_{\tilde{\mathbf{a}}} \frac{\left(\mathbf{Y}-\mathbf{X}\mathbf{H}_G^{-1}\tilde{\mathbf{a}}\right)^T \mathbf{H}_W^T\mathbf{H}_W\left(\mathbf{Y}-\mathbf{X}\mathbf{H}_G^{-1}\tilde{\mathbf{a}}\right)}{1+\tilde{\mathbf{a}}^{\mathbf{T}}\tilde{\mathbf{a}}}. \quad (14)$$

where $\mathbf{W}=\mathbf{H}_W^T\mathbf{H}_W$.

The weighted biased normal system of equations is defined as

$$\left(\left(\mathbf{X}\mathbf{H}_G^{-1}\right)^{\mathbf{T}}\mathbf{W}\mathbf{X}\mathbf{H}_G^{-1}-\sigma_W^2\mathbf{I}\right)\tilde{\mathbf{a}}=\left(\mathbf{X}\mathbf{H}_G^{-1}\right)^T\mathbf{W}\mathbf{Y}, \quad (15)$$

where $\sigma_W=\sigma_{\min}\left(\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1},\mathbf{H}_W\mathbf{Y}\right)$ is smallest singular value of a matrix $\left(\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1},\mathbf{H}_W\mathbf{Y}\right)$.

The system (13) is often ill-conditioned. The condition number of a normal biased system is determined from the following expression

$$cond_2\left(\left(\mathbf{X}\mathbf{H}_G^{-1}\right)^{\mathbf{T}}\mathbf{W}\mathbf{X}\mathbf{H}_G^{-1}-\sigma_W^2\mathbf{I}\right)=\frac{\sigma_{\max}^2\left(\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1}\right)-\sigma_W^2}{\sigma_{\min}^2\left(\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1}\right)-\sigma_W^2}.$$

Ill-conditioning (15) arises for two reasons: because of the product $\left(\mathbf{X}\mathbf{H}_G^{-1}\right)^{\mathbf{T}}\mathbf{W}\mathbf{X}\mathbf{H}_G^{-1}$ and because of the possible proximity of $\sigma_{\min}^2\left(\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1}\right)$ and σ_W^2

To increase the stability of the solution of the biased normal system of equations, the Cholesky decomposition can be applied. Cholesky's decomposition has a significant drawback - on ill-conditioned matrices, it can lead to an unacceptable error in solving the system of equations.

Performing transformations similar to [20] we get the augmented system equivalent to weighted biased normal system of equations (15) is defined as

$$\overline{X}\overline{\tilde{a}}=\overline{Y}, \quad (16)$$

or

$$\left(\begin{array}{c|c|c} \mathbf{I} & \mathbf{0} & k_W\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1} \\ \hline \mathbf{0} & \mathbf{I} & jk_W\sigma_W\mathbf{I} \\ \hline k_W\left(\mathbf{X}\mathbf{H}_G^{-1}\right)^T\mathbf{H}_W^T & jk_W\sigma_W\mathbf{I} & \mathbf{0} \end{array}\right)\left(\begin{array}{c} k_W\bar{\mathbf{e}} \\ \hline k_W\mathbf{r} \\ \hline \hat{\tilde{\mathbf{a}}} \end{array}\right)=\left(\begin{array}{c} k_W\mathbf{H}_W^T\mathbf{Y} \\ \hline \mathbf{0} \\ \hline \mathbf{0} \end{array}\right). \quad (17)$$

The condition number of the matrix $\overline{X}(k)$ is determined by the expression [18]:

$$cond_2\left(\bar{X}(k)\right)=\frac{\sqrt{1+k^2\mu_{\max}+k^2\sigma_W^2}}{\sqrt{1+k^2\mu_{\min}+k^2\sigma_W^2}}\times$$
$$\times\left|\frac{\cos\left(\frac{1}{3}\arccos\left(\frac{3\sqrt{3}}{2}\frac{k^2\left(\mu_{\max}-\sigma_W^2\right)}{\left(1+k^2\mu_{\max}+k^2\sigma_W^2\right)^{3/2}}\right)\right)}{\cos\left(\frac{\pi}{3}+\frac{1}{3}\arccos\left(\frac{3\sqrt{3}}{2}\frac{k^2\left(\mu_{\min}-\sigma_W^2\right)}{\left(1+k^2\mu_{\min}+k^2\sigma_W^2\right)^{3/2}}\right)\right)}\right|, \quad (18)$$

where $\mu_{\max}, \mu_{\min}$ are the maximum and minimum eigenvalues of the matrix $\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1}\left(\mathbf{X}\mathbf{H}_G^{-1}\right)^T\mathbf{H}_W^T$.

The problem of finding the minimum value of the condition number can be considered as the problem of choosing the optimal factor:

$$\min_{k>0} cond_2\left(\bar{\mathbf{X}}(k)\right) \quad (19)$$

Equation (19) has no analytical solution, but can be solved by numerical methods. In practice, the value can be used as an estimate

$$\hat{k}_W=\frac{\sigma_{\max}\left(\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1}\right)+\sigma_W}{\sigma_{\min}\left(\mathbf{H}_{\mathbf{W}}\mathbf{X}\mathbf{H}_G^{-1}\right)+\sigma_W}\sqrt{\frac{2}{\sigma_{\max}^2\left(\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1}\right)+\sigma_W^2}}. \quad (20)$$

Since the true values of the matrices are unknown, the weighted total least squares algorithm can be implemented iteratively.

Algorithm 1.

Step 1. Find matrices $\hat{\mathbf{W}}$ и $\hat{\mathbf{G}}$ and the decompositions $\hat{\mathbf{W}}=\hat{\mathbf{H}}_W^T\hat{\mathbf{H}}_W$, $\hat{\mathbf{G}}=\hat{\mathbf{H}}_G^T\hat{\mathbf{H}}_G$ of the matrices. A new vector of variables introduces (13). The initial values of the matrices are $\hat{\mathbf{W}}=\mathbf{I}$,

$$\mathbf{G}=\begin{pmatrix}1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & 1/2\end{pmatrix}.$$

Step.2. Find the smallest singular value $\sigma_W=\sigma_{\min}\left(\mathbf{H}_W\mathbf{X}\mathbf{H}_G^{-1},\mathbf{H}_W\mathbf{Y}\right)$.

Step 3. Calculate the value of the multiplier $k_W>0$ by (20).

Step 4. Solve the system of equations (17) using one of the standard methods, for example, LU decomposition [19].

Step 5. Repeat **Step1-Step4** until a reliable solution has been reached.

Step 6. Find the parameter estimate $\hat{\mathbf{a}}_{WTLS}=\mathbf{H}_G^{-1}\hat{\hat{\mathbf{a}}}$.

V. Numerical Results

In section we present numerical examples which illustrated the performance proposed algorithm. The numerical results were obtained using Matlab.

The noise-free signal model is described by the equation:

$$s(n)=\sum_{l=1}^{5}\alpha_l\sin\left(\omega_l n+\varphi_l\right), \quad (21)$$

where $\boldsymbol{\alpha}=\begin{pmatrix}1 & 1 & 1 & 1 & 1\end{pmatrix}$, $\boldsymbol{\omega}=\begin{pmatrix}1.35 & 1.70 & 2.05 & 2.4 & 2.75\end{pmatrix}$, $\boldsymbol{\varphi}=\begin{pmatrix}\pi & \pi/2 & \pi/3 & \pi/4 & \pi/5\end{pmatrix}$.

Signal-to-noise ratio for the sum of sinusoids is [25]

$$SNR=10\lg\left(\frac{\alpha_1^2}{2\sigma_\xi^2}\right). \quad (22)$$

The frequency estimates obtained using the proposed algorithm 1 were compared with the Yule-Walker estimates (YW [11]), total least square (TLS [8]), total least square for the model with symmetric coefficients (8), (TLS-S), least square for the model with symmetric coefficients (8), (LS-S), Cramer-Rao lower bound (CRLB).

The algorithms were compared by misalignment of the coefficient estimation vector $\hat{a}$:

$$\delta\omega=20\lg\left(\frac{\left\|\hat{\boldsymbol{\omega}}-\boldsymbol{\omega}\right\|_2}{\left\|\boldsymbol{\omega}\right\|_2}\right). \quad (23)$$

The results of numerical experiments for different N are presented in Fig. 1-3.The results presented in this section show that the proposed algorithm allows one to obtain estimates for small samples and high SNR values.

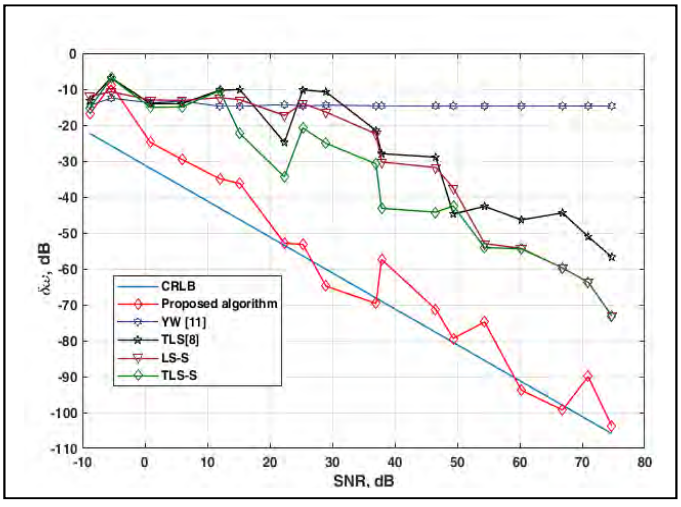


Fig. 1. Misalignment of the coefficient estimation vector $\hat{a}$ for $N=25$.

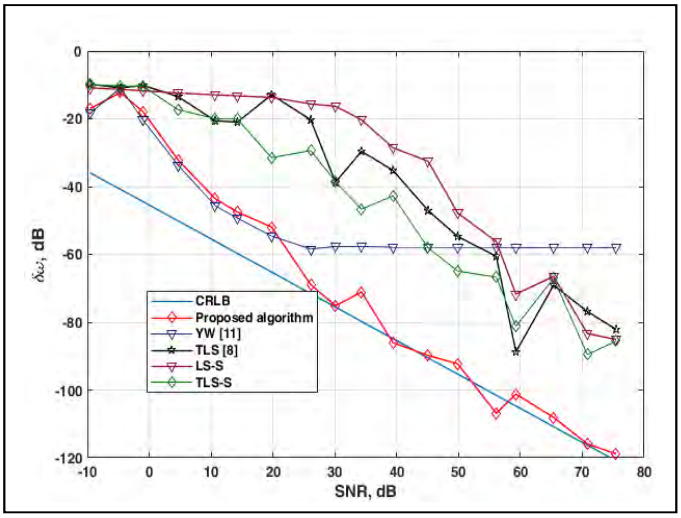


Fig. 2. Misalignment of the coefficient estimation vector $\hat{a}$ for $N = 75$.

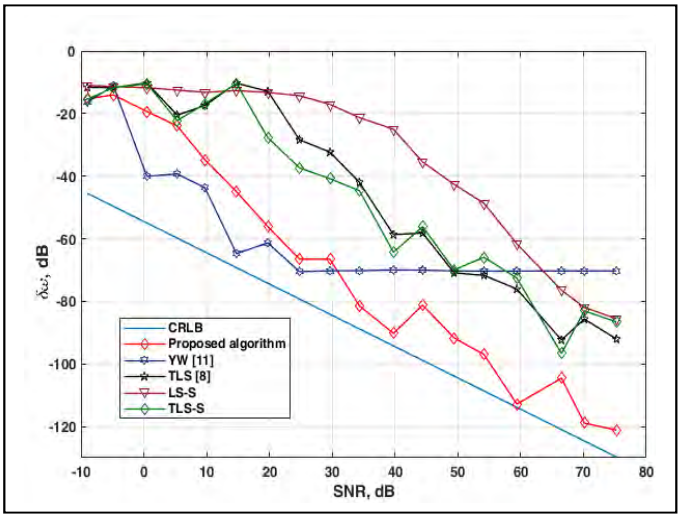


Fig. 3. Misalignment of the coefficient estimation vector $\hat{a}$ for $N = 150$.

VI. CONCLUSION

The paper proposes a method for estimating the frequencies of sinusoids with noise based on the solution of a weighted biased normal system. The ability to obtain a solution to the total least squares problem without finding the right singular vector and matrix is $\left(\mathbf{X}\mathbf{H}_G^{-1}\right)^T \mathbf{W}\mathbf{X}\mathbf{H}_G^{-1} - \sigma_W^2\mathbf{I}$ an undoubted advantage of the frequency estimates based on augmented systems, compared to solutions based on the singular value decomposition of a matrix or a biased normal system, respectively.

REFERENCES

[1] S.L. Marple., Digital Spectral Analysis with Applications. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[2] P. Stoica, R. Moses, Introduction to Spectral Analysis. Upper Saddle River, NJ: Prentice-Hall, 1997.

[3] P. Stoica P., “List of references on spectral line analysis”, Signal Process. vol. 31, no. 3, 1993, pp. 1298–1319.

[4] S.M. Kay, Modern Spectral Estimation : Theory and Application. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[5] B. G. Quinn and E. J. Hannan, The Estimation and Tracking of Frequency Cambridge, U.K.: Cambridge Univ. Press, 2001.

[6] D.C. Rife and R.R. Boorstyn, “Multiple tone parameter estimation from discrete-time observations”, Bell Syst. Tech. J. 1976. pp. 1389–1410.

[7] P. Stoica and A. Nehorai, “Statistical analysis of two nonlinear least squares estimators of sine wave parameters in the colored noise case,” in Proc. Int. Conf. Acoust. Speech, Signal Processing, vol. 4, New York, 1988, pp. 2408–2411.

[8] M. D. Rahman and K. B. Yu, “Total least squares approach for frequency estimation using linear prediction,” IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-35, no. 5, pp. 1440–1454, Oct. 1987.

[9] D.V..Ivanov, O.A. Katsyuba, B.K. Grigorovskiy, “Determination of frequency in three-phase electric circuits with autocorrelated noise”, Russ. Electr. Engin. vol. 88. 2017. pp. 123–126 .

[10] Y. T. Chan, J. M. M. Lavoie, and J. B. Plant, “A parameter estimation approach to estimation of frequencies of sinusoids,” IEEE Trans. Acoust. Speech, Signal Process., vol. ASSP-29, no. 2, pp. 214–219, Apr. 1981.

[11] Y. T. Chan and R. P. Langford, “Spectral estimation via the high-order Yule-Walker equations,” IEEE Trans. Acoust. Speech, Signal Processing, vol. ASSP-30, no. 5, pp. 689–698, Oct. 1982.

[12] Li T.H., Kedem B. “Iterative filtering for multiple frequency estimation”, IEEE Trans. Signal Processing. vol. 42, no. 5. 1994. pp. 1120–1131.

[13] Y. Hua and T. K. Sarkar, "On the total least squares linear prediction method for frequency estimation," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, no. 12, pp. 2186-2189, Dec. 1990, doi: 10.1109/29.61547.

[14] S. H. Leung, T. H. Lee and W. H. Lau, "Total Least SquaresLinear Prediction for Frequency Estimation with FrequencyWeighting," Proc. IEEE, ICASSP,Munich, Germany, vol. 5, pp. 3993-3996, Apr. 1997.

[15] P. Stoica, T. Söderström, S. Van Huffel, “On SVD-based and TLS-based high-order Yule-Walker methods of frequency estimation”,Signal Processing, vol. 29, no. 3.1992. pp.309-317.

[16] P. Stoica, T. Söderström, “Comparative performance study of SVD-based and QRD-based high-order Yule-Walker methods for frequency estimation”, Circuits Systems and Signal Process, vol. 12. 1993. pp. 105–117.

[17] H. So, K.W. Chan, Y. Chan, Y., & K.C. Ho, “Linear prediction approach for efficient frequency estimation of multiple real sinusoids: algorithms and analyses”. IEEE Transactions on Signal Processing, vol. 53. 2005. pp.2290-2305.

[18] J.H.Wilkinson, The Algebraic Eigenvalue Problem. Oxford University Press, Inc. 1988.

[19] I. Markovsky, S. Van Huffel, “Overview of total least-squares methods” Signal Processing. vol.87, no.10. 2007. pp. 2283-2302.

[20] A.I. Zhdanov, P.A. Shamarov, “Direct projection method in the problem of complete least squares”, Autom. Remote Control. 2000. vol.. 61, no. 4. 2020. pp. 610-620.

[21] D.V. Ivanov et al, “Identification of exponential trend models with fractional white noise” , J. Phys.: Conf. Ser., 2019. 1368 042061.

[22] W.F. Trench, “Inversion of Toeplitz band matrices”. Mathematics of computation, vol. 28 no. 128. 1974, pp. 1089-1095

[23] G. Ammar, W. Gragg , “Superfast solution of real positive definite Toeplitz systems”,b SIAM J. Matrix Anal. Appl. vol. 9. 1988. pp. 61–76.

[24] G.H Golub ; van Loan C.F. Matrix Computations (3rd ed.), Johns Hopkins University Press, 1996.

[25] P. Stoica , R.L. Moses , T. Soderstrom , J. Li , “Optimal high-order Yule-Walker estimation of sinusoidal frequencies”, IEEE Transactions on Signal Processing. vol. 39, no. 6. 1991. pp.

Detection of motor imagery (MI) event in Electroencephalogram (EEG) signals using artificial intelligence technique

Muhammad Yeamim Hossain
Department of Electronics and Telecommunication Engineering
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
yeamim.sumon@gmail.com

A. B. M. S. U. Doulah
Department of Electrical and Electronic Engineering
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
abul.sayeed@ulab.edu.bd (*Corresponding author*)

***Abstract*— With the recent development of technology and acquisition devices, the research of detection and classification utilizing EEG signals is rapidly increasing. One of the critical research in the field of the brain-computer interface includes an accurate detection of motor neuron behavior called motor imagery (MI) events. Due to the increased number in people with inabilities (e.g. paralyzed people, autism, and elderly people), accurate detection of MI events can of great help. In this work, a method for the detection of the MI events using the electroencephalogram (EEG) signal is proposed. Data from thirteen random subjects from a publicly available dataset was utilized. Firstly, the EEG signals were preprocessed and then a combination of time domain and frequency domain features were extracted from the signals. The number of features was reduced and selected using a minimum-redundancy-maximum-relevance (MRMR) algorithm and forward feature selection. On the subject level with leave-one-subject-out cross-validation, MI events were recognized with an average F1-score of 68.69% using the Support Vector Machine classification model. The best individual performance was obtained with an F1-score of 79%. These results suggest that the proposed approach is able to identify MI events in the EEG signal and thus the method may potentially be integrated into devices that can assist people with inability. Further improvement in the performance of the method can be done by carrying out testing in a wider population.**

Keywords—Electroencephalograph (EEG), motor imagery (MI), movement, Support Vector Machine (SVM), prediction, classification

I. INTRODUCTION

Advances in electroencephalogram (EEG) signal processing and sophisticated computing capabilities have potential possibilities to diagnose underlying diseases in humans. One of the uses of the EEG signal is to understand and diagnose the motor imagery (MI) events, the mental processing of imagining movements without incurring any actual physical movements [1]. Specifically, the motor imagery information can be useful for the paralyzed/elderly people who have the inability to perform physical movements at their old age and/or during rehabilitation. Therefore, it is important to analyze the MI events in EEG signals in order to explore potential solutions to assist physically handicapped and/or elderly people.

Recent efforts are focused on analyzing and studying MI events signals and trying to understand the signals in-depth in order to develop potential systems that can assist treating patients or the people who are unable to express themselves [2] - [3]. Different groups of patients under different clinical conditions are studied using motor imagery EEG signals towards solving the underlying issues. Several studies worked on MI events to study stroke patients and explored whether the motor recovery can be gained through mental practice or not. It was observed that the MI events can be helpful to activate the brain motor area. But the acute stroke patients may not be benefitted from mental practice [2]. Neurofeedback can help patients to learn the MI strategy and can increase the MI effectiveness to improve the health conditions [4]. The study of MI events was also shown to be very useful to learn a new sport more efficiently through mental practice in combination with physical practice [3]. The work in [3] showed that imagery rehearsal can be very effective in the improvement of learning a new sport and it was far better than the only physical practice or without practice. Notable research works observed that the event-related motor imagery EEG signals can be suggested to bring improvement in the area of rehabilitation. Motor imagery movement can also be helpful to assist physically handicapped patients. The MI based movement can be implemented instead of physical movement to reduce the problems of expressing the needs of such individuals.

With a goal of developing an accurate MI detection system, numerous researchers are trying to improve the algorithms to get better and efficient recognition. In [5], the authors showed the advantages of the improved back propagation (BP) neural network over the traditional BP neural network in MI event detection by solving the low signal-to-noise ratio and unclean filtering issues. The work presented in [6] proposed an optimized motor imagery paradigm where significant improvement was found in classification accuracy and usability. The imagery data of the hand movement of the writing pattern of a Chinese character was used as motor imagery data. Then the common spatial pattern (CSP) method was utilized to extract the features and support vector machine (SVM) was used for the classification. The authors in [7] proposed a method to detect the motor imagery movement using linear discriminant analysis (LDA) classifier. In [8], the

work proposed an empirical mode decomposition (EMD) based method to detect the mu rhythm during motor imagery hand movement. In a recent study [9], the authors used a method to detect motor imagery left and right hand movement using support vector machine (SVM) classifier where independent component analysis (ICA) was used to remove noise signals of motor disable person. For control and stroke rehabilitation, EEG based strategies are also studied to detect MI events. The study of [10] used the filter bank common spatial pattern (FBCSP) algorithm to decompose the EEG into multiple frequency pass bands and later on the common spatial pattern algorithm was used to extract the features for the band pass frequency ranges. Despite the existing work on MI events detections, the accuracy of the detection is yet to be improved. Due to the dynamic nature of the EEG signal and dependency of the subject's compliance, further improvement in the method of MI event detection is critical. In the present work, a method to predict the MI events from the movement of the hand and different fingers is proposed. The work also demonstrates how feature selection algorithm can affect the recognition accuracy. To improve accuracy, the best feature set from the feature selection algorithm is finalized. Finally, several key points from the results are discussed in detail.

II. Materials and Methods

A. Data Acquisition

The raw motor imagery EEG data were obtained from publicly available "GigaDB" database [11]. The dataset contains EEG recordings of both hands MI tasks. The data were acquired using 64 channel Ag/AgCl active electrodes at the sampling rate of 512 Hz. A total of 13 subjects were considered randomly from the dataset for this study. Apart from MI EEG data, the database also includes non-MI tasks such as eye blinking, eyeball movement, etc. However, only the MI event data were considered for the study. A summary of the data protocol is presented in Fig. 1.

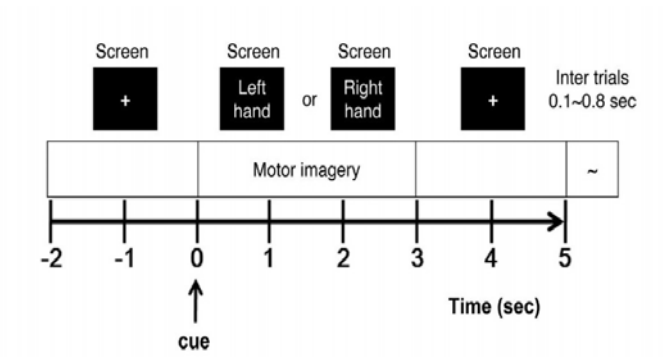


Fig. 1. Setup for each trial [11]

A total of 100 trials were taken for MI data. Each trial duration was 7s where a black screen appeared for the first 2s, motor imagery activity occurred for 3s and a black screen reappeared for the next 2s (shown in Fig. 1). Therefore, the total duration of 100 trials was 700s. A total of sixty-four electrodes were used to form motor imagery data and each electrode consisted of imagery data of 700 seconds. The data was organized into MI data from left hand movement and right hand movement as MI-left and MI-right respectively [11]. In the present work, only MI-left was utilized for the development of the classification model and validation.

B. Data Processing and Annotation

Firstly, the bad trials were identified by using the band-passed filter, and later a high-pass filter above 0.5 Hz was used to remove drifts from all EEG trials. Next, the data was divided into consecutive non-overlapping frames. The raw EEG data for 3 electrodes obtained from the MI tasks is illustrated in Fig.2. The dynamic nature of EEG signals is evident from the electrode data. Next, the data annotation was performed to mark the MI events and the period of non-MI events and/or idle. The onset of the MI events was first spotted and then for a period of 3s was marked as MI events. The rest of the data in one trial was assumed as non-MI events. Since the annotation was done by human raters, the annotation reliability was assessed with the kappa coefficient. A good kappa coefficient of 0.80 suggested a reliable annotation.

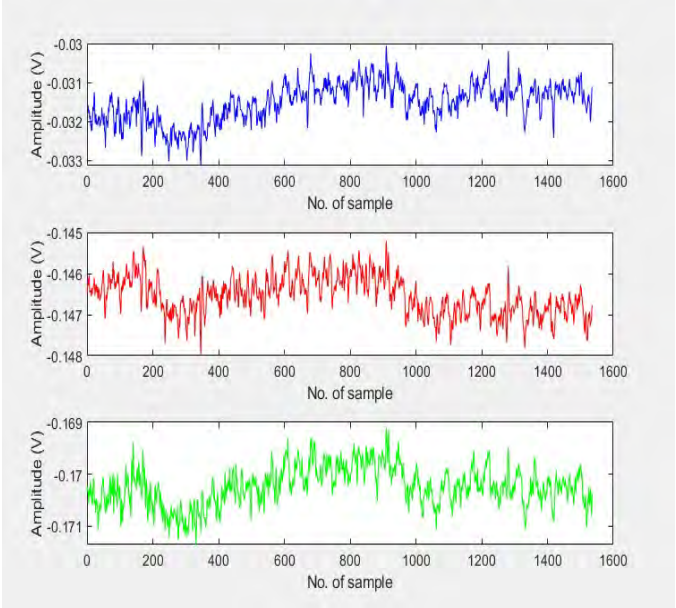


Fig. 2. Raw EEG data from MI tasks

C. Feature Extraction and Selection

Feature extraction and selection is an important step towards analyzing EEG data. A total of seventeen features are extracted from each of the non-overlapping epochs for each electrode which contains 700s of sensor data. Table I. presents the list of features that are used in this study. In order to eliminate any possible dimensional inconsistencies, the feature vector was normalized in the range of -1 to 1.

TABLE I. LIST OF FEATURES

No.	Features	No. of Electrode	Total features
1	Mean Value	64	17 × 64 = 1088
2	Median Value		
3	Standard Deviation		
4	Mean Absolute Deviation		
5	Quantile25		
6	Quantile75		
7	Signal Interquartile Range		
8	Sample Skewness		
9	Sample Kurtosis		
10	Spectral Entropy		
11	Peak2Peak Value		
12	RMS Value		
13	Crest Factor		
14	Shape Factor		
15	Impulse Factor		
16	Margin Factor		
17	Signal Energy		

To minimize the cost and complexity of models, the feature set was reduced using feature selection procedures. Out of 13 subjects, 3 subjects were utilized for feature selection. To select the most important features, the minimum redundancy maximum relevant (MRMR) algorithm [12] was applied to rank the features sequentially based on mutual information. Different numbers of top features were checked (i.e. "k" values 10, 20, 30, and 40 where the k is the number of features that will be ranked sequentially to check). After ranking the features, Forward feature selection (FFS) was applied to find out the best sequence of the features from the ranked features. Linear discriminant analysis (LDA) was used during the forward feature selection process to find out the features in a sequence.

D. Classification

Different machine learning algorithms perform differently, based on the application of data sets. The Support Vector Machine (SVM) is one of the most important tools for machine learning and widely implemented in data classification. Out of several advantages, good generalization and regularization are two major powerful properties that outperform other classifiers. Due to these properties, SVM was utilized to detect and classify MI events. A linear kernel function was computed in this study for its simplicity and less complexity. MATLAB classification learner package is used to train and test the model. To test the performance of the classification model leave one subject out procedure was used to find out the performance of each subject where among the 10 subjects, 9 subjects were used to train the model and 1 subject was used to test the performance. This procedure was followed for all 10 subjects. Therefore, 10 training models were generated and the label for the 10 test data was predicted respectively. Finally, the predicted label of all 10 subjects was found.

E. Performance Evaluation

To validate the performance of the proposed method for the detection of the MI events, different performance metrics were measured. The parameters were positive (P) that defined epoch was MI event, negative (N) that defined epoch was non-MI event, true positive (TP) that defines epoch was positive and predicted to be positive, true negative (TN) that defines epoch was negative and predicted to be negative, false positive (FP) that defined observation was positive and predicted to be negative, and false negative (FN) that defined epoch was negative predicted to be positive [16]. The result was calculated by using the parameter from the data where it measured accuracy, sensitivity, specificity, precision, recall, and F1-score. The equations for the performance metrics were as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{P} \quad (2)$$

$$Specificity = \frac{TN}{N} \quad (3)$$

$$\Pr ecision = \frac{TP}{TP + FP} \quad (4)$$

$$F - score = \frac{2TP}{2TP + FP + FN} \quad (5)$$

III. RESULTS

For different numbers of ranking features in MRMR, the best performance was obtained when the top 10 features where ranked by the MRMR algorithm. After the FFS, a total of 8 features were selected from the feature selection algorithm. The classification results for the different numbers of MRMR features (i.e. 10, 20, 30, and 40) in terms of the accuracy, sensitivity, specificity, precision, and F1-score are reported in Tables II -V. The mean values and the standard deviation are also given mentioned respectively.

It is evident from Table II that the six subjects show accuracy above 70% where 5 subjects show their F1-score value above 70%. The average accuracy and F1-score scores were 68.29% and 68.69% respectively. Table III shows that the average accuracy and F1-score scores were 69.01% and 68.55% respectively. In Table IV, the average accuracy and F1-score score were found to be 68.43% and 68.14% respectively. Finally, Table V exhibited that the average accuracy and F1-score score were found at 66.56% and 66.72% respectively.

TABLE II. PERFORMANCE OF THE SUBJECTS WITH MRMR FEATURE VALUE 10

Sub No.	Acc	Sens	Specs	Prec.	F1-score
1	71.00%	75.33%	67.75%	63.66%	69.01%
2	56.14%	95.67%	26.50%	49.40%	65.15%
3	53.29%	83.00%	31.00%	47.43%	60.36%
4	75.43%	92.00%	63.00%	65.09%	76.24%
5	82.14%	82.33%	82.00%	77.43%	79.81%
6	72.57%	81.33%	66.00%	64.21%	71.76%
7	77.71%	84.67%	72.50%	69.78%	76.51%
8	70.57%	81.00%	62.75%	61.99%	70.23%
9	59.43%	50.33%	66.25%	52.80%	51.54%
10	64.57%	81.33%	52.00%	55.96%	66.30%
Avg	68.29%	80.70%	58.98%	60.78%	68.69%
SD	9.10	11.52	16.76	8.90	7.97

TABLE III. PERFORMANCE OF THE SUBJECTS WITH MRMR FEATURE VALUE 20

Sub No.	Acc	Sens	Specs	Prec	F1-score
1	70.71%	74.67%	67.75%	63.46%	68.61%
2	59.29%	82.33%	42.00%	51.57%	63.41%
3	53.71%	78.33%	35.25%	47.57%	59.19%
4	75.86%	94.00%	62.25%	65.13%	76.94%
5	81.29%	82.33%	80.50%	76.00%	79.04%
6	72.71%	81.33%	66.25%	64.38%	71.87%
7	77.57%	84.00%	72.75%	69.81%	76.25%
8	72.43%	82.33%	65.00%	63.82%	71.91%
9	60.29%	48.33%	69.25%	54.10%	51.06%
10	66.29%	80.67%	55.50%	57.62%	67.22%
Avg	69.01%	78.83%	61.65%	61.35%	68.55%
SD	8.43	11.19	13.14	8.20	8.27

TABLE IV. PERFORMANCE OF THE SUBJECTS WITH MRMR FEATURE VALUE 30

Sub No.	Acc	Sens	Specs	Prec	F1-score
1	71.00%	76.67%	66.75%	63.36%	69.38%
2	57.14%	80.67%	39.50%	50.00%	61.73%
3	55.14%	86.67%	31.50%	48.69%	62.35%
4	69.43%	77.33%	63.50%	61.38%	68.44%
5	80.57%	84.00%	78.00%	74.12%	78.75%
6	72.29%	86.67%	61.50%	62.80%	72.83%
7	75.71%	86.33%	67.75%	66.75%	75.29%
8	73.29%	78.33%	69.50%	65.83%	71.54%
9	62.14%	44.00%	75.75%	57.64%	49.91%
10	67.57%	93.67%	48.00%	57.46%	71.23%
Avg	68.43%	79.43%	60.18%	60.80%	68.14%
SD	7.69	12.83	14.69	7.30	7.84

TABLE V. PERFORMANCE OF THE SUBJECTS WITH MRMR FEATURE VALUE 40

Sub No.	Acc	Sens	Specs	Prec	F1-score
1	66.14%	78.00%	57.25%	57.78%	66.38%
2	57.00%	79.33%	40.25%	49.90%	61.26%
3	53.00%	81.67%	31.50%	47.21%	59.83%
4	71.14%	77.67%	66.25%	63.32%	69.76%
5	72.14%	83.00%	64.00%	63.36%	71.86%
6	72.29%	85.33%	62.50%	63.05%	72.52%
7	77.43%	84.67%	72.00%	69.40%	76.28%
8	70.57%	84.67%	60.00%	61.35%	71.15%
9	60.71%	48.67%	69.75%	54.68%	51.50%
10	65.14%	81.33%	53.00%	56.48%	66.67%
Avg	66.56%	78.43%	57.65%	58.65%	66.72%
SD	7.29	10.26	12.26	6.44	7.00

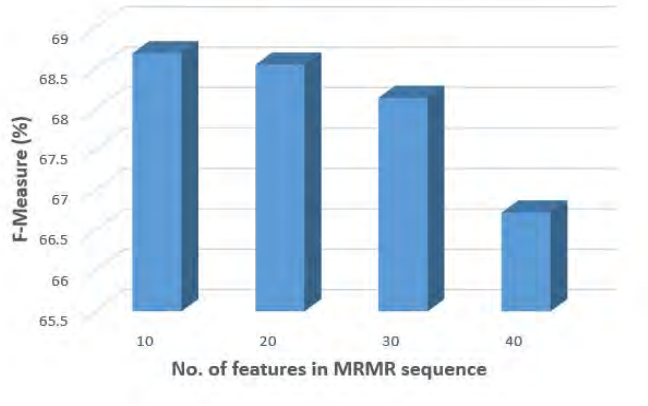


Fig. 3. Average F1-score score with different no. of MRMR sequence

It was evident that the best average F1-score of 68.59 was in case of taking 10 top features at the MRMR mutual information ranking stage. It was also evident from the results that the standard deviation of the F1-score tends to decrease as the increase of MRMR features. Fig. 3 illustrates a comparison in the F1-score with the different number of features in the MRMR sequence.

IV. DISCUSSION

The aim of the current work was twofold: i) to develop a method of the detection of the MI events in EEG signal and ii) identify the best set of feature and how the number of feature set affect the performance of the model. To the end, the investigation led to a method that can detect the MI-events almost 70% correctly. Also, the work provided an analysis of how the number of features can affect performance.

It is to be noted that the signal strength of EEG is very low and at some point, it goes below the threshold of noise. Therefore, it was extremely difficult to process the signal with such dynamic behavior. However, the current method could

mostly remove the noise and was able to process the signal towards the detection of the MI-events.

The best results were found to be on an average F1-score of around 70% which may be suitable for brain-computer interface with limited applications. It was also evident that some of the subjects exhibited an F1-score of almost close to 80% which leads to a positive direction of further improving the performance. One of the interesting results from the classification model is the narrow standard deviation. It is always expected to have a narrow standard deviation in accuracy provided that the models don't get over fitted.

Another contribution of the current work was that the analysis of the number of features initially selected using mutual information. From the performance of the subjects with MRMR feature value 10, 20, 30, and 40 it was observed that the results are pretty much around the corner. It was evident that the best performance was obtained when the MRMR algorithm selected the top 10 ranked features. As the number of features increased, the performance of the model started to fall off. One potential reason could be the quality of the extracted features. There could be features those are highly correlated to each which in turn contributes to error. The best performance was obtained using only 8 features. This small number could potentially help to direct real-time analysis.

The method presented in the paper considers non-overlapping epochs of 512 samples. To obtain optimal performance, overlapping epochs can be considered in future studies. Since MI-events are dynamic, there are possibilities that tiny epochs would perform better compared to a big window.

While the method presents accurate detection of MI-events, the study was not free from limitations. One of the major limitations was that the method was to validate on a wider range of populations. The number of subjects can be increased to test and validate the model. Future work should be done considering more participants with more variety in MI events. Different classifier's performance can be explored to detect the MI-events. Further analysis can be carried out to detect which electrodes affect the performance of the classifier the most. This finding may lead to process fewer number of electrodes which eventually decrease the computational complexity.

V. Conclusion

In this study, a method for detecting motor imagery (MI) events in EEG signals was introduced using machine learning algorithm. On an average of ~70% F1-score was found from 10 subjects trials. The minimum redundancy maximum relevant (MRMR) algorithm followed by the forward feature selection (FFS) reduced the feature set drastically which helps to gain computational time. The initial results could potentially be helpful to develop a brain-computer interface that can contribute to help people with disabilities such as paralyzed people, elderly people, autistic people, etc. Of course, the accuracy can be improved further, however, the method provides a positive direction toward developing a more accurate algorithm. Further works can be done using more participants, different domains feature, and different classifiers.

References

[1] M. Jeannerod, "Mental imagery in the motor context," *Neuropsychologia*, vol. 33, no. 11, pp. 1419–1432, Nov. 1995, doi: 10.1016/0028-3932(95)00073-C.

[2] M. Ietswaart *et al.*, "Mental practice with motor imagery in stroke recovery: randomized controlled trial of efficacy," *Brain J. Neurol.*, vol. 134, no. Pt 5, pp. 1373–1386, May 2011, doi: 10.1093/brain/awr077.

[3] C. Frank, W. M. Land, C. Popp, and T. Schack, "Mental Representation and Mental Practice: Experimental Investigation on the Functional Links between Motor Memory and Motor Imagery," *PLOS ONE*, vol. 9, no. 4, p. e95175, Apr. 2014, doi: 10.1371/journal.pone.0095175.

[4] C. Kranczioch, C. Zich, I. Schierholz, and A. Sterr, "Mobile EEG and its potential to promote the theory and application of imagery-based motor rehabilitation," *Int. J. Psychophysiol.*, vol. 91, no. 1, pp. 10–15, Jan. 2014, doi: 10.1016/j.ijpsycho.2013.10.004.

[5] L. Liu, "Recognition and Analysis of Motor Imagery EEG Signal Based on Improved BP Neural Network," *IEEE Access*, vol. 7, pp. 47794–47803, 2019, doi: 10.1109/ACCESS.2019.2910191.

[6] Z. Q iu *et al.*, "Optimized Motor Imagery Paradigm Based on Imagining Chinese Characters Writing Movement," *IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc.*, vol. 25, no. 7, pp. 1009–1017, 2017, doi: 10.1109/TNSRE.2017.2655542.

[7] P. H orki *et al.*, "Detection of mental imagery and attempted movements in patients with disorders of consciousness using EEG," *Front. Hum. Neurosci.*, vol. 8, p. 1009, 2014, doi: 10.3389/fnhum.2014.01009.

[8] Guo Xiaojing, Wu Xiaopei, and Zhang Dexiang, "Motor imagery EEG detection by empirical mode decomposition," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008, pp. 2619–2622, doi: 10.1109/IJCNN.2008.4634164.

[9] A. Dey, S. Bhattacharjee, and D. Samanta, "Recognition of motor imagery left and right hand movement using EEG," in *2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, May 2016, pp. 426–430, doi: 10.1109/RTEICT.2016.7807856.

[10] K. K. Ang and C. Guan, "EEG-Based Strategies to Detect Motor Imagery for Control and Rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 4, pp. 392–401, Apr. 2017, doi: 10.1109/TNSRE.2016.2646763.

[11] H. Cho, M. Ahn, S. Ahn, M. Kwon, and S. C. Jun, "EEG datasets for motor imagery brain-computer interface," *GigaScience*, vol. 6, no. 7, pp. 1–8, 01 2017, doi: 10.1093/gigascience/gix034.

[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[13] Yujun Yang, Jianping Li, and Yimei Yang, "The research of the fast SVM classifier method," in *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Dec. 2015, pp. 121–124, doi: 10.1109/ICCWAMTIP.2015.7493959.

[14] M. I. Hejazi and X. Cai, "Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm," *Adv. Water Resour.*, vol. 32, no. 4, pp. 582–593, Apr. 2009, doi: 10.1016/j.advwatres.2009.01.009.

[15] D. Ververidis and C. Kotropoulos, "Sequential forward feature selection with low computational cost," in *2005 13th European Signal Processing Conference*, Sep. 2005, pp. 1–4.

[16] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behav. Processes*, vol. 148, pp. 56–62, Mar. 2018, doi: 10.1016/j.beproc.2018.01.004.

[17] A. Tharwat, "Classification assessment methods," *Appl. Comput. Inform.*, Aug. 2018, doi: 10.1016/j.aci.2018.08.003.

[18] C. Zhang *et al.*, "Feature selection for high dimensional imbalanced class data based on F1-score optimization," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, Dec. 2017, pp. 278–283, doi: 10.1109/SPAC.2017.8304290.

SOI Instrumentation Amplifier for High-Temperature Applications

Evgenii V. Balashov
Higher school of applied physics and space technologies
Peter the Great St.Petersburg Polytechnic University
St.-Petersburg
balashov_ev@mail.ru

Nikita V. Ivanov
Higher school of applied physics and space technologies
Peter the Great St.Petersburg Polytechnic University
St.-Petersburg
ivanovnick@mail.ru

Alexander S. Korotkov
Higher school of applied physics and space technologies
Peter the Great St.Petersburg Polytechnic University
St.-Petersburg
korotkov@spbstu.ru

***Abstract*— The article presents the results of the development and measurements of a specialized integrated circuit of an instrumentation amplifier (IA) with indirect negative current feedback. The schematic and the measurement results of the main characteristics for a silicon-on-insulator metal-oxide-semiconductor instrumentation amplifier are presented. The amplifier is intended for use as a part of a sensor network for monitoring the state of high-temperature objects. The temperature range of the instrumentation amplifier is up to 200 °C. Instrumentation amplifier has 11.8 MHz gain bandwidth product with a 3.3 V supply voltage. The measured gain is 27.50 dB and the error in setting the gain was 0.27 dB, the 3 dB bandwidth is 500 kHz.**

Keywords*— SOI, metal oxide semiconductor, instrumentation amplifier, high temperature

I. INTRODUCTION

An urgent problem is the development of a high-temperature electronic component base for the system of object monitoring and controlling. The system usually includes a sensor and an interface part for preliminary processing of the signal taken from the sensor and its digitization. In the interface part, the instrumentation amplifier is most critical component since it determines the dynamic range, sensitivity, and interferer immunity of the system. Especially important nowadays is the monitoring system for high temperature application e.g. for engine monitoring system. To ensure the high-temperature operating mode of the instrumentation amplifier (up to 200 °C), it is good solution to use the silicon-on-insulator (SOI) technology [1-4].

The article presents the results of the development and measurements of a specialized integrated circuit of an instrumentation amplifier (IA) with indirect negative current feedback. Instrumentation amplifiers are widely used in the processing of signals from various sensors to amplify signals with high accuracy. In this case, the amplifier is intended for use as part of a sensor network for monitoring the state of high-temperature objects.

The schematic of the instrumentation amplifier is described in the second section after the short introduction. The measurement results are presented in the third section. The conclusion is made at the end of the article.

II. SCHEMATIC OF INSTRUMENTATION AMPLIFIER

A. Architecture of the Instrumention Amplifier

The most widespread are two main circuitry solutions of an instrumentation amplifier. The first one is the instrumentation amplifier based on three operational amplifiers. A block diagram of the first solution, an instrumentation amplifier based on three operational amplifiers, is shown in Figure 1. This circuit is traditional for instrumentation amplifiers based on discrete components. To eliminate the influence of the operational amplifier gain mismatch, the instrumentation amplifier gain is completely determined by resistive feedback circuit. The instrumentation amplifier consists of two non-inverting amplifiers and a differential amplifier. Non-inverting amplifiers are based on operational amplifiers A_1 and A_2 with a feedback circuit on resistors R_2, R_1 and R_3, R_1, respectively. Non-inverting amplifiers provide high input impedance, gain and load capacity sufficient to minimize the effect of the finite input impedance of a differential amplifier based on an operational amplifier A_3 with feedback circuits on resistors R_4, R_5, R_6, R_7. The differential amplifier provides the calculation of the difference between the input signals and the gain. If the equality

$$R_7/R_6 = R_5/R_4$$

is fulfilled, the signal from both inputs of the differential amplifier is transmitted to the output with the equal. Thus, taking into account the gain of the non-inverting amplifiers at the input, the gain of the instrumentation amplifier is defined as

$$G = \frac{R_5}{R_4}\left(\frac{R_2 + R_3}{R_1} + 1\right).$$

The second solution is an amplifier with indirect current feedback [5-11], shown in Figure 3. This architecture is used if instrumentation amplifier circuit is integrated on a single chip. In this case, the gain is determined not only by the passive

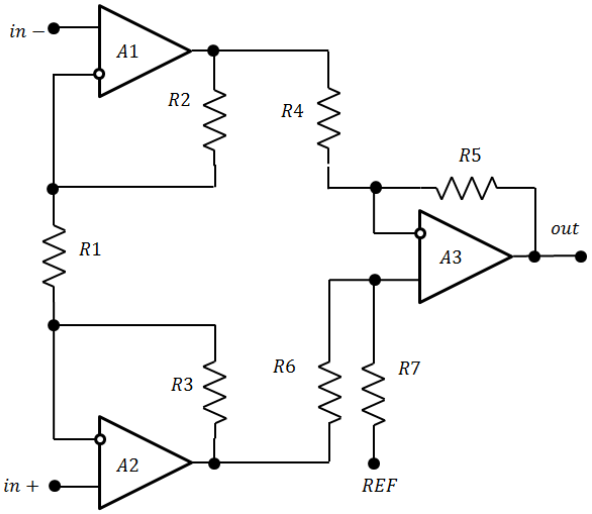


Fig. 1. Block diagram of an instrumentation amplifier.

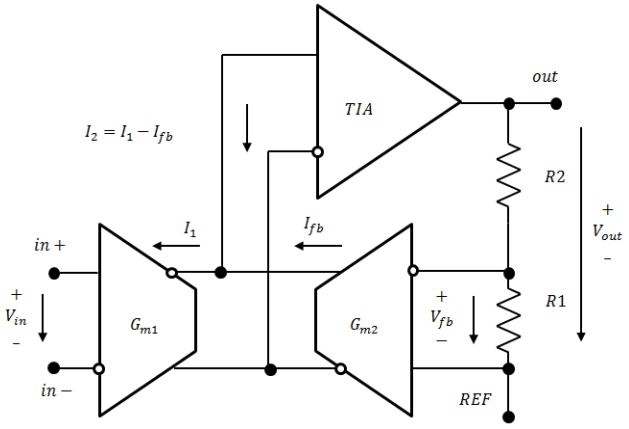


Fig. 2. Block diagram of an instrumentation amplifier.

feedback loop, but also by the gains of the transconductance amplifiers with gains G_{m1} and G_{m2}. The transconductance amplifier can be well matched. Thus, at $G_{m1} = G_{m2}$, the gain (A_V) of the amplifier depends only on the resistive feedback circuit and is determined by the following expression

$$A_V = 1+R_2/R_1.$$

The operation principle of this amplifier is based on the fact that initially the input signal V_{in} is converted into a current I_1, which in turn is converted by a transimpedance amplifier (TIA) with a transresistance into an output voltage V_{out} relative to the reference voltage V_{ref}. The output of the amplifier is connected to a feedback circuit in the form of a resistive divider across resistors R_1 and R_2, which converts the output voltage V_{out} into a voltage V_{fb}. The voltage V_{fb} is applied to the input of a transconductance amplifier with a transfer conductance G_{m2} and is converted into a current I_{fb}, which is subtracted from the

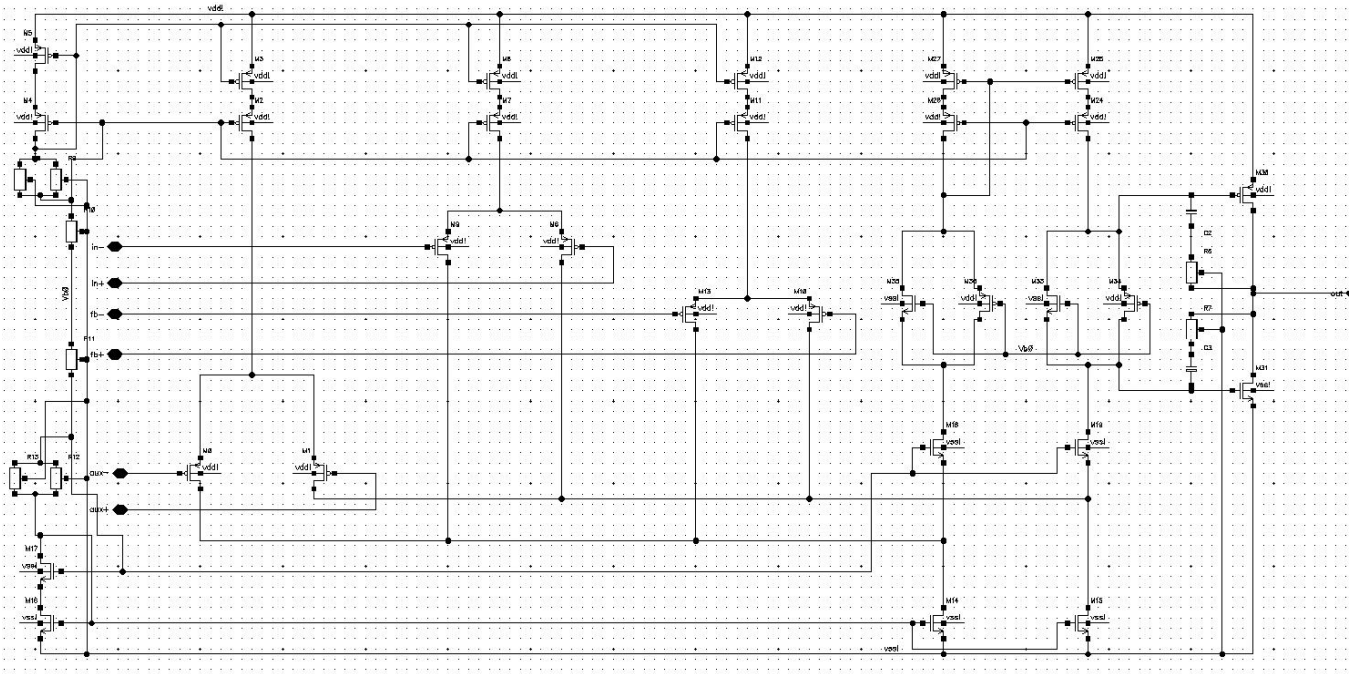


Fig. 3. Schematic of an instrumentation amplifier.

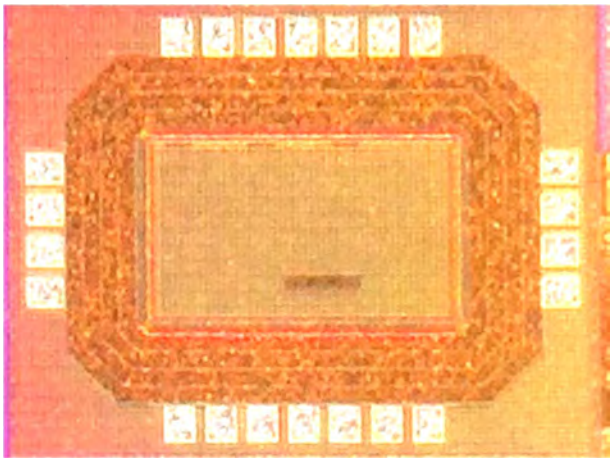


Fig. 4. Block diagram of an instrumentation amplifier.

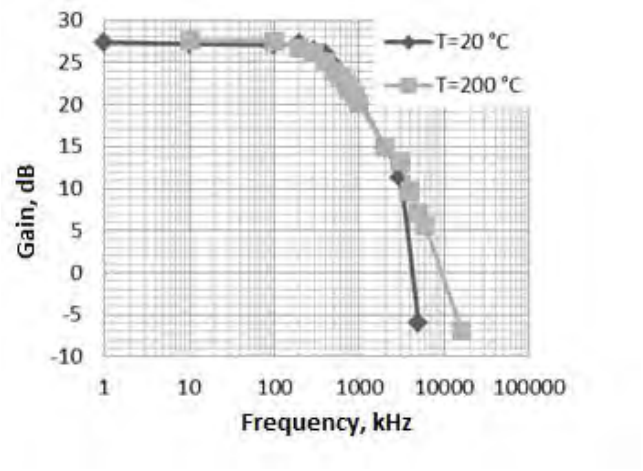


Fig. 5. Block diagram of an instrumentation amplifier.

current I_1, closing the feedback loop. This circuit includes the minimum number of passive elements that can be off-chip implemented to provide the required gain.

B. Schematic of the Instrumention Amplifier

A schematic of instrumentation amplifier with the indirect current feedback is shown in Figure 3 without external off chip feedback loop. The input signal goes to a pair of input transistors, that form the first differential amplifier. The first amplifier converts the input signal to current form. The cascoded current source is used to increase the common mode reject ratio of the differential amplifier. The second differential amplifier is used to convert feedback signal to current form. The currents from the differential amplifiers are summarized at the input of transimpedance amplifier that is formed by common gate amplifier and push pull amplifier as an output stage. The third differential transistor pair is used to compensate offset voltage. The phase margin is provided by an RC circuit.

The integrated circuit is designed and manufactured using XT018 silicon-on-insulator technology (HV SOI 0.18 um CMOS) with a minimum resolution of 0.18 microns provided by X-FAB. The dimensions of the instrumentation amplifier crystal, taking into account the ESD protection ring and contact pads, were 1.026x0.826 mm (0.85 mm2). A photograph of the integrated circuit crystal of the developed amplifier is shown in Figure 4.

III. Measurement Results

To carry out the measurements, the instrumentation amplifier chip was soldered onto a test printed circuit board, on which power filtering circuits and discrete gain setting resistors R1 and R2 were placed. An instrumentation amplifier with indirect current feedback according to the results of an experimental study has the following parameters. The supply voltage is 3.3 V. The current consumption in the absence of an input signal was less than 1 mA, which corresponds to a power consumption in static mode 3.3 mW. The power consumption in

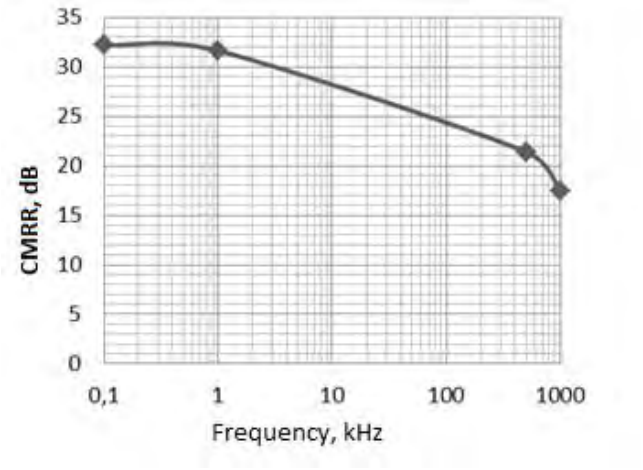


Fig. 6. Block diagram of an instrumentation amplifier.

dynamic mode is up to 9.9 mW, which is explained by the push-pull circuit of the amplifier's output stage.

The gain measurements of the instrumentation amplifier were carried out at a nominal value of 27.23 dB, which is determined by resistors R2 and R1 with a nominal value of 330 and 15 Ohms respectively. Resistors with a tolerance of 1% from the nominal value were used, which leads to a possible error in the gain from 25.57 dB to 28.91 dB. The measured gain was 27.50 dB. Thus, the gain error is determined by the technological deviation of discrete resistors.

The amplifier zero offset V_{offset} was measured at unity gain by measuring the output voltage at zero differential input signal. To do this, we set a unity gain with a resisistor R_1 of 1.5 MΩ with a reference resistance R_2 of 15 Ω. The Vin + and Vin- inputs

are short-circuited and provide a constant common-mode bias voltage V_{cm} of 1.65 V. At supply voltage V_{DD} = 3.3 V and reference voltage V_{ref} = 1.65 V, the level of the DC voltage component at the output V_{out} is measured on the output. The DC deviation from V_{ref} represents the offset voltages $V_{offset} = V_{out} - V_{ref}$. The measured offset voltage at unity gain is less than 10 mV.

To measure the frequency response, the V_{in+}, V_{in-} inputs are supplied with a differential input signal and a DC in-phase component from an Agilent 81150A waveform generator operating in differential mode. The results of measurements of the frequency response of the instrumentation amplifier in the operating frequency band at a nominal gain of 27.50 dB are shown in Figure 5, and at room temperature (T = 20 ° C) and at elevated temperature (T = 200 ° C). When the operating temperature changes from 20 to 200 degrees Celsius, the gain changes in the operating frequency band do not exceed 0.4 dB. The 3 dB bandwidth was 500 kHz, so the gain area is 11.8 MHz. The measurements were made at temperature 200 degrees. A slight decrease in gain is observed (no more than 1.2 dB at a maximum gain of 27 dB). For unity gain, no deviation of the gain from the nominal value was recorded.

The common-mode signal rejection ratio was measured by changing the input common-mode component from 1.65 V to 2.4 V, i.e. by 750 mV, while the DC component at the amplifier output is changed less than 1 mV. Thus, the CMR is greater than 57 dB. According to the simulation results, the CMRR was 75 dB. The frequency dependence of the common-mode reject ratio (CMRR) is shown in Figure 6. Measurements were made with a nominal gain of 27.50 dB. The common-mode reject ratio at low frequencies is 32.4 dB and 21.4 dB at 500 kHz.

To determine the operating range of the input common-mode signal of the amplifier at unity gain, the constant common-mode voltage at the input was changed from 0 V to a supply voltage of 3.3 V. The gain did not change up to the input common-mode bias voltage of 2.4 V. When this voltage is exceeded, the transistors of the input differential stage of the amplifier go into cut-off. Thus, the input common mode range is from 0 V to 2.4 V.

The instantaneous values of the output signal voltages are in the range from 0.4 to 2.8 V, which gives us an output voltage swing of 2.4 V. When the output swing exceeds more than 2.4 V, the amplifier output signal goes into saturation.

At maximum gain, the lower limit of the allowable supply voltage was determined. The shape and amplitude of the signal did not change when the supply voltage dropped to 3 V. Thus, this amplifier operates at 3 V or more.

IV. Conclusion

The instrumentation amplifier is developed using indirect current feedback architecture using XT018 technology from XFAB with a minimum resolution of 0.18 um. The dimensions of the chip of the amplifier with indirect current feedback, taking into account the electrostatic protection ring and contact pads, were 1.026 x 0.826 mm (0.85 mm sq).

The obtained measurement results confirm the operability of the integrated circuit of the SOI MOS instrumentation amplifier at temperatures up to 200 degrees Celsius. The parameters of the amplifier with indirect current feedback were measured. The supply voltage is from 3 V to 3.6 V. The power consumption in static mode is no more than 3.3 mW and in dynamic mode up to 9.9 mW. The gain bandwidth product is more than 6.6 MHz, input common-mode signal range from 0 to 2.4 V, maximum voltage swing of the output signal is 2.4 V; the coefficient of the common-mode signal rejection is not less than 57 dB.

References

[1] P. C. de Jong and G. C. M. Meijer, "A high-temperature electronic system for pressure-transducers," in IEEE Transactions on Instrumentation and Measurement, vol. 49, no. 2, pp. 365-370, April 2000.

[2] P. C. de Jong G. C. M. Meijer and A. H. M. van Roermund " A 300 °C dynamic-feedback instrumentation amplifier" IEEE J. Solid-State Circuits vol. 33 pp. 1999-2009 Dec. 1998.

[3] J. Pathrose, C. Liu, K. T. C. Chai, Y. Ping Xu, "A Time-Domain Band-Gap Temperature Sensor in SOI CMOS for High-Temperature Applications," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 62, no. 5, May 2015, pp. 436–440.

[4] M. Malits, I. Brouk, Y. Nemirovsky, "Temperature sensing circuits in CMOS-SOI technology," 2017 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS), Tel-Aviv, 2017, pp. 1–5.

[5] B. J. van den Dool and J. H. Huijsing, "Indirect Current Feedback Instrumentation Amplifier with a Common Mode Input Range That Includes the Negative Rail," ESSCIRC '92: Eighteenth European Solid-State Circuits conference, Copenhagen, 1992, pp. 175-178.

[6] Bernard van den Dool, and Johan Huijsing, "Indirect Current-Feedback Instrumentation amplifier with a common-mode input range that includes the negative rail," IEEE Journal of Solid-State Circuits, vol. 28, no. 7, July 1993, pp. 743–749.

[7] R. Wu, K. A. A. Makinwa and J. H. Huijsing, "A Chopper Current-Feedback Instrumentation Amplifier With a 1 mHz $1/f$ Noise Corner and an AC-Coupled Ripple Reduction Loop," in IEEE Journal of Solid-State Circuits, vol. 44, no. 12, pp. 3232-3243, Dec. 2009.

[8] M. A. P. Pertijs and W. J. Kindt, "A 140 dB-CMRR Current-Feedback Instrumentation Amplifier Employing Ping-Pong Auto-Zeroing and Chopping," in IEEE Journal of Solid-State Circuits, vol. 45, no. 10, pp. 2044-2056, Oct. 2010

[9] A. Catania, S. Del Cesta, P. Bruschi and M. Piotto, "Design of current feedback instrumentation amplifiers with rail-to-rail input-output ranges," 2017 13th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME), Giardini Naxos, 2017, pp. 125-128.

[10] Jiongming Wang, Fuding Ge, Shengqi Yang, Xinnan Lin and Jin He, "Low gain-error instrumentation amplifier for current sensing," 2010 IEEE International Conference of Electron Devices and Solid-State Circuits (EDSSC), Hong Kong, 2010, pp. 1-4.

[11] J. M. Carrillo, M. A. Domínguez, R. Pérez-Aloe, J. F. Duque-Carrillo and C. A. de la Cruz, "CMOS Low-Voltage Indirect Current Feedback Instrumentation Amplifiers With Improved Performance," 2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Genoa, Italy, 2019, pp. 262-265.

Bit-Stream Power Function Online Computer

A.S. Shkil
Design Automation Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine,
oleksandr.shkil@nure.ua

L.V. Larchenko
Design Automation Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine,
lina.larchenko@nure.ua

B.D. Larchenko
Design Automation Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine,
bogdan.larchenko@gmail.com

***Abstract*— The article proposes an approach to the bit-stream power function online computer design based on the forming increments method of increasing step functions, which provides the principle of sampling a certain part of the bits from the input bit stream, which allows input bit stream frequency increase that ensures the minimal computation time. In this case, the absolute error in calculating the given function does not exceed half of the argument's least significant bit unit. The device has a streaming method for online computations with parallel-sequential conversions over the input bit stream in accordance with a given function. The hardware implementation of the device is based on the Moore finite state machine, which increases the device's reliability. The results of modeling a bit-stream power function online computer's behavioral model are presented. The device is implemented in the FPGA platform, which provides flexibility of reconfiguration, reliability and high performance.**

Keywords*— *functional conversion, bit-stream data, bit-stream computing, approximation, mathematical model, VHDL, verification, finite state machine

I. INTRODUCTION

Currently, there is an increase in the complexity of tasks for organizing computations in sensor systems, robotics, control systems and intelligent measuring systems. One of the directions associated with the creation of new basic elements for the construction of these systems is the development of devices that perform functional transformation of information signals represented by bit data streams. [1, 2].

In systems where primary processing of measurement information takes place to make decisions about measurement results in order to implement control tasks, in most cases, smooth changes in control signals are required, for example, when acting on the actuators of robots, manipulators, and other devices. At the same time, various functions can be used to smooth the signals: logarithmic, exponential, power, trigonometric [3].

Most digital systems work with a positional representation of data, such as binary coding. An alternative is to represent the data as a bit sequence in a specific time interval. This representation is much less compact than binary encoding, but complex operations can be performed using simple logic [4]. In bit stream coding, data are streams of single amplitude pulses. In this case, an informative parameter is a fixed value of arbitrary duration pulses for a time interval.

Bit-stream signal forms allow the transmission and processing of information by methods characterized by the possibility of sequential streams processing at the rate of a single bits arrival and high noise immunity due to the non-positional nature and equivalence of single bits in relation to their weight in a digital code [5]. At the same time, the bit-stream form of signals, while maintaining noise immunity, does not provide information redundancy and allows for high performance of devices.

Bit-stream power function online computes are widely used in measuring and computing technology, where there is a digital functional sweep, which provides for the implementation of the streaming method of computations in time with the simultaneous parallel-sequential execution of transformations on single bits of the data stream in accordance with the required function, i.e. sequential computation of function values performed for adjacent argument values. Each subsequent value of the function is calculated based on the previous calculation result. In this case, the first calculation is carried out taking into account additional information (input of initial conditions) [6].

When synthesizing bit-stream power-law computers with a fractional exponent, an approach is used in which the procedure of calculating the function is carried out in two stages, in the first stage, the numerical value of the x^m in a parallel code is calculated, and in the second, the root of the n-th degree. In this case, the intermediate function x^m is formed by the m-stage inclusion of binary multipliers, the output frequency of each one of them increases from stage to stage, which limits the frequency of the input bit stream.

The proposed bit-stream power function online computer allows combining the exponentiation and root extraction operations in one device. This allows to increase the input bit stream frequency due to the generating unit increments of the step function method use, which is based on the selection of certain bit numbers from the input bit stream corresponding to the approximation nodes of the power function. On the basis of the obtained bit-stream power function computer's mathematical model, the hardware implementation of the device was made. The hardware implementation is performed by building a model using the hardware description language in a synthesized VHDL subset and subsequent synthesis using the Xilinx CAD tools. The FPGA platform was chosen as an effective element base for the implementation of the device, which provides the flexibility of reconfiguration, high speed and technological reliability of the device which in the fututre promises the possibility of massively parallel computing at relatively good power efficiency[7].

II. MATHEMATICAL MODEL OF BIT-STREAM POWER FUNCTION COMPUTER

The mathematical model of the bit-stream power function computer was obtained on the basis of increasing step functions increments forming method [8]. According to

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

the method, the continuous function $y^* = f(x^*)$, limited by conditions $x^*, y^* \geq 0$, $y^* \leq x^*$, $\frac{dy^*}{dx^*} > 0$ and having an inverse function $x^* = \psi(y^*)$, can be reproduced in the output of a bit-stream computer by a step approximating function of the following form

$$y = [f(x) + |\delta_{max}|], \quad (1)$$

where x, y - input and output bit-stream data, respectively, $|\delta_{max}|$ – specified boundary value of the continuous functions reproduction absolute error, lying in the range $0.5 \leq |\delta_{max}| < 1$. In (1), square brackets denote the integer part of a number

The process of reproducing function (1) can be reduced to sampling a certain part of the bits from the input bit-stream data x, the numbers of which are determined by the inequality

$$\Psi(y - |\delta_{max}|) \leq x_y < \Psi(y - |\delta_{max}|) + 1, \quad (2)$$

where $\Psi(y - |\delta_{max}|)$ – is a inverse function of $f(x)$.

Inequality (2) is a formula for the general term x_y in the numerical sequence $x_1, x_2, x_3, \ldots$, of selected bits from the input bit-stream x, which form the output bit-stream y and correspond to the nodes of approximation of the power function. In this case, the values of x_y, can be found by successive substitution of the y = 1, 2, 3,... in inequality (2), calculating its left side and rounding the obtained discrete values upward to the nearest integer.

Bit-stream power function online computer is designed to calculate continuous functions of the form

$$y^* = x^{*\frac{m}{n}}, \quad (3)$$

where m, n – are positive natural numbers.

The power step function approximating a continuous (3) has the form

$$y = [x^{\frac{m}{n}} + |\delta_{max}|]. \quad (4)$$

The formation of the power step function (4) at the computer output can be carried out by the simultaneous formation of function increments x^m and y^n in the process of entering the bit-stream x at the input of the device, continuous comparison of their current values and the formation of the output bits y at the moments of their equality.

Provided that $m < n$ the values of the power function samples x_y can be determined based on the expression (2)

$$(y - |\delta_{max}|)^{\frac{n}{m}} \leq x_y < (y - |\delta_{max}|)^{\frac{n}{m}} + 1. \quad (5)$$

When substituting in (5) the values of the minimal absolute error $|\delta_{max}| = 0.5 = \frac{1}{2}$ after some transformations, an inequality was obtained

$$(2y-1)^n \leq x_y^m 2^n < (2y-1)^n + 1. \quad (6)$$

The inequality implemented in the device based on (6) has the form

$$2^n x_y^m \geq (2y-1)^n. \quad (7)$$

The bit-stream power function online computer mathematical model is a system of inequalities, which is obtained on the basis of inequality (7) and written in the differences

$$\begin{aligned} &2^n x_1^m \geq (2y_1 - 1)^n, \\ &2^n (x_2^m - x_1^m) + \Delta_1 \geq (2y_2 - 1)^n - (2y_1 - 1)^n, \quad (8) \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &2^n (x_y^m - x_{y-1}^m) + \Delta_{y-1} \geq (2y_k - 1)^n - (2y_{k-1} - 1)^n, \end{aligned}$$

where Δ_{y-1} – the difference obtained as a result of comparing the increments of the current values of the functions $2^n x^m$ and $(2y-1)^n$ between two adjacent nodes of the approximation of the reproduced step function; integer values $y_1 \leq y \leq y_k$ and $1 \leq y_k \leq k$.

In the system of inequalities (8) Δ_1 and Δ_{y-1} is defined as

$$\Delta_1 = 2^n x_1^m - (2y_1 - 1)^n, \quad (9)$$

$$\Delta_{y-1} = 2(x_y^m - x_{y-1}^m) + \Delta_{y-2} - (2y_k - 1)^n + (2y_{k-1} - 1)^n. \quad (10)$$

When a bit x_y of a bit-stream x arrives at the input of the device, an output bit y_k will be generated at its output when each inequality of system (8) is satisfied. In this case, the first bit out $y_1 = 1$ corresponds to the selected bit with the number x_1 of the input bit-stream x and the first inequality of system (8) will be satisfied. Similarly, the second bit $y_2 = 2$ corresponds to the bit with the number x_2, at which the second inequality of the system will be satisfied, and so on.

As an example, consider a power function

$$y = [x^{\frac{2}{3}} + 0.5], \quad (11)$$

where the specified absolute computation error is $|\delta_{max}| = 0.5$.

Based on (7), the inequality that must be implemented in the device has the form

$$2^3 x_y^2 \geq (2y-1)^3. \quad (12)$$

Using formula (12), the bit-stream power function computer mathematical model (11) takes the form

$$2^3 x_1^2 \geq (2y_1 - 1)^3,$$

$$2^3(x_2^2 - x_1^2) + \Delta_1 \geq (2y_2 - 1)^3 - (2y_1 - 1)^3, \quad (13)$$

. . . . . .

$$2^3(x_y^2 - x_{y-1}^2) + \Delta_{y-1} \geq (2y_k - 1)^3 - (2y_{k-1} - 1)^3.$$

In the system of inequalities (13) Δ_1 and Δ_{y-1} is defined as

$$\Delta_1 = 2^3 x_1^2 - (2y_1 - 1)^3, \quad (14)$$

$$\Delta_{y-1} = 2^3(x_y^2 - x_{y-1}^2) + \Delta_{y-2} - (2y_k - 1)^3 + (2y_{k-1} - 1)^3. \quad (15)$$

III. Hardware Implementation of Bit-Stream Power Function Computer

The hardware implementation of bit-stream power function online computer is performed using the example of function (11).

TABLE I shows the results of calculating the values of function (11) with a given calculation error $|\delta_{max}| = 0.5$ and rounding of the function result for values of the input bit-stream of length $x_{max} = 10$.

TABLE I. Results of Calculating the Power Function

X	Y
x = 1	$y = [1^{\frac{2}{3}} + 0.5] = [1.5] = 1$
x = 2	$y = [2^{\frac{2}{3}} + 0.5] = [2.09] = 2$
x = 3	$y = [3^{\frac{2}{3}} + 0.5] = [2.58] = 2$
x = 4	$y = [4^{\frac{2}{3}} + 0.5] = [3.02] = 3$
x = 5	$y = [5^{\frac{2}{3}} + 0.5] = [3.42] = 3$
x = 6	$y = [6^{\frac{2}{3}} + 0.5] = [3.8] = 3$
x = 7	$y = [7^{\frac{2}{3}} + 0.5] = [4.16] = 4$
x = 8	$y = [8^{\frac{2}{3}} + 0.5] = [4.5] = 4$
x = 9	$y = [9^{\frac{2}{3}} + 0.5] = [4.83] = 4$
x = 10	$y = [10^{\frac{2}{3}} + 0.5] = [5.14] = 5$

Sample values x_y from the input bit-stream are determined based on the expression (5)

$$x_y = [(y - 0.5)^{\frac{3}{2}}] + 1. \quad (16)$$

When substituting the values y = 1, 2, 3, 4, 5 in (16), the values of the samples x_y were obtained: 1, 2, 4, 7, 10 respectively.

The device implements the inequality (12). The calculation of the left and right sides of inequality (12) can be carried out on the basis of pipeline calculations, which imply the calculation of arithmetic series of the 2nd and 3rd orders, respectively, by substituting the values x, y = 1, 2, 3, ..., 10. In this case, the issues of device synthesis are solved by lowering the order of the arithmetic series differences.

For the $2^3 x_y^2$ function, the second-order arithmetic series is

8, 32, 72, 128, 200, 288, 392, 512, 648, 800.

Arithmetic series of differences of the first and second orders are

Δ 8, 24, 40, 56, 72, 88, 104, 120, 136, 152

Δ^2 16, 16 , 16, 16, 16, 16, 16, 16, 16.

For the function $(2y - 1)^3$ the third-order arithmetic series is

1, 27, 125, 343, 729, 1331, 2197, 3375, 4913, 6859.

Arithmetic series of differences of the first, second and third orders, respectively, are

Δ 26, 98, 218, 386, 602, 866, 1178, 1538, 1946

Δ^2 72, 120, 168, 216, 264, 312, 360, 408

Δ^3 48, 48, 48, 48, 48, 48, 48.

The considered technique is used to construct a pipelined architecture of a bit-stream calculator of a power function, which contains polynomial modules consisting of the following components: adder SUM_1, summing counter Count, register RG1, which implement the function $2^3 x_y^2$ and adders SUM_1, SUM 2, SUM _3, register RG2, which implement the function. $(2y - 1)^3$. In the adder SUM_1, the parallel codes of the lattice function $2^3 x_y^2$ increments are compared with the lattice function $(2y - 1)^3$ increments, taking into account their difference obtained at the previous stage of calculations. The increment of the lattice function $2^3 x_y^2$ enters SUM_1 as direct binary code, and the increment of the lattice function $(2y - 1)^3$ – as additional code.

The components of the device architecture are initialized with the initial values of the obtained arithmetic series of the second and third orders and their differences, respectively: Count = 8, SUM1 = -1, SUM2 = 26, SUM3 = 72, RG1 = 16, RG2 = 48.

TABLE II shows the computational process in the device components. The results of evaluating the function (TABLE I) and the appearance of the overflow bits y at the output of the device are the same.

TABLE II. Computing Process in Device Components

X	SUM_1	Y	Count	SUM_2	SUM_3
1	–1+8=7	1	8+16=24	26+72=98	72+48=120
	7–26= –19				
2	–19+24 =5	1	24+16=40	98+120=218	120+48=168
	5–98=-93				
3	–93+40 = –53		40+16=56		
4	–53+56=3	1	56+16=72	218+168=386	168+48=216
	3–218= -215				
5	–215+72= –143		72+16=88		
6	–143+88= –55		88+16=104		
7	–55+104= 49	1	104+16=120	386+216=602	216+48=264
	49–386 = –337				
8	-337+120=–217		120+16=136		
9	–217+136= –81		136+16=152		
10	–81 +152 = 71	1	152+16=168	602+264=866	264+48=312
	71–602= –531				

When 10 bits of the bit-stream x are fed to the input of the device, 5 bits of the output bit-stream y will appear at its

output. Bits with numbers 1, 2, 4, 7, 10 will be selected from the input bit-stream and fed to the device output, which is confirmed by the calculated values of the x_y samples.

When performing hardware implementation, a block diagram of the designed device was developed. The device contains two units: a pulse detector and a bit-stream calculator unit (Fig. 1).

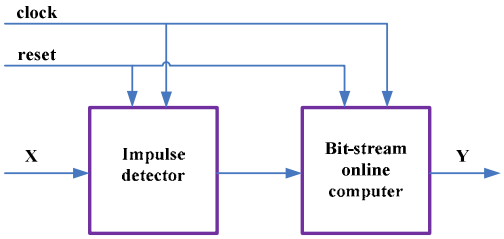


Fig. 1. Block diagram of the device

The impulse detector unit is designed to detect the bits of the input bit stream x and at the output sets the corresponding impulse = 1 signal, which will be received by the arithmetic unit of the computer for further processing.

The block of the bit-stream online calculator contains the arithmetic block "Powerfunc", which performs the operation of cocking the argument x to a fractional power with a given absolute computation error. The arithmetic block issues a Ready signal, which means that this block is ready to receive the next bit for processing. The result of the block operation is the output bit-stream signal y, which is the result of calculating the power function.

Bit-stream power function online computer is implemented on the basis of a Moore machine and is represented by a composition of operational and control units. A graph-scheme of the algorithm was developed in accordance with the mathematical model of the calculator, which describes the principle of device operations (Fig. 2).

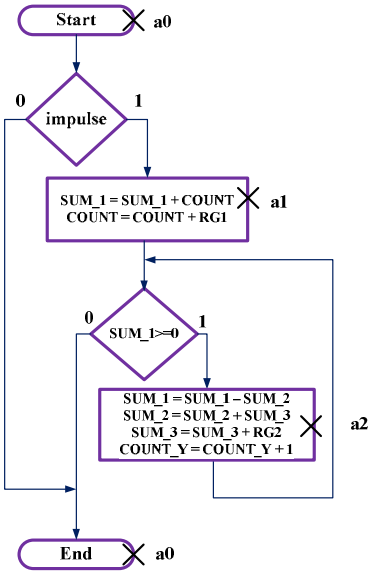


Fig. 2. Graph-diagram of the device operation algorithm

The device works according to the following algorithm.

When a reset signal is applied (reset = 1), the device registers are set to the values: Count = 8, SUM1 = -1, SUM2 = 26, SUM3 = 72, RG1 = 16, RG2 = 48, count_Y = 0.

When the next bit arrives at the input of the device, the value of the SUM1 register is increased by the value of the Count register, and the value of the Count register is increased by the value of the RG1 register.

If the value of the SUM1 register is positive, then the output bit is generated at the device output, the value of Count_Y is increased by one, the value of the SUM2 register is subtracted from the value of the SUM1 register, the value of the SUM2 register is added to the value of the SUM3 register, the value of the SUM3 register is increased by the value of the RG2 register.

Point 3 is repeated as long as the value of the register SUM1 is positive.

The control unit is described by the transition graph, which is obtained as a result of the synthesis of the algorithm graph-scheme for the Moore machine. The transition graph has 3 states: a0, a1, a2 (Fig. 3).

By the reset = 1 signal, the unit switches to the initial state a0 and remains in this state until the impulse signal appears from the output of the pulse detector. With the arrival of the impulse signal, the unit goes into state a1.

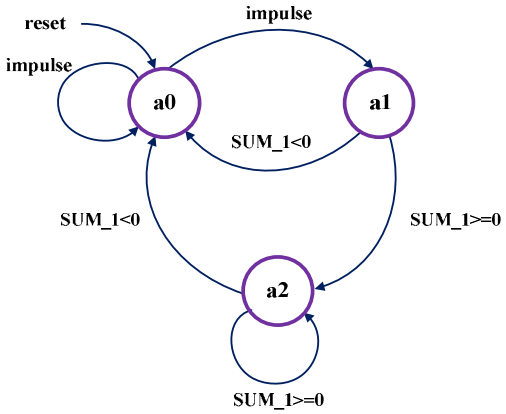


Fig. 3. The transition graph of the arithmetic block control unit

In state a1, if $SUM1 \geq 0$, the unit will go into state a2, if $SUM1 < 0$, the unit will go into state a0.

The unit stays in state a2 if $SUM1 \geq 0$. If the value of the $SUM1 < 0$, the unit will go to state a0. The unit also issues a signal to generate the output bit y of the device.

The Active-HDL software package was chosen as an effective environment for modeling and verification of projects on the FPGA platform, which allows automating the process of entering a project into a CAD system. The hardware model of the device is developed in the VHDL language using an automatic description model.

Fig. 4 shows a fragment of the program code describing operational unit work process, which shows the initialization of the components Count, SUM1, SUM2, SUM3. This code implements serial-parallel computations in the arithmetic unit of the device.

```
process (clock_i, reset_i)
begin
if (reset_i = '1') then
counter <= CONV_STD_LOGIC_VECTOR(8,width1);
sum_1 <= CONV_STD_LOGIC_VECTOR(-1, width2);
sum_2 <= CONV_STD_LOGIC_VECTOR(26, width2);
sum_3 <= CONV_STD_LOGIC_VECTOR(72, width1);
else
        if (falling_edge(clock_i)) then
                if (sum_plus_a_i = '1') then
                sum_1 <= sum_1 + counter;
                counter <= counter + 16;
                else
                if (sum_minus_b_i = '1') then
                count <= count + 1;
                sum_1 <= sum_1 - sum_2;
                sum_2 <= sum_2 + sum_3;
                sum_3 <= sum_3 + 48;
                end if;
                end if;
        end if;
end if;
end process;
```

Fig. 4. Fragment of the program code describing operational unit computing process

Fig. 5 contains a program code fragment, describing the control unit operations in accordance with the transition graph.

The Fig. 6 shows a waveform, with the results of modeling the behavioral model of a bit-stream power function online computer, which shows that the values in the register of the SUM1 component coincide with the calculated data of the computing process and the appearance of the output bits of the device y corresponds to the sample numbers x_y. The Fig. 7 shows RTL–scheme of synthesized device.

```
begin
case (state) is
        when a_0 =>
        if x_i = '1' then next_state <= a_1;
        else next_state <= a_0;
        end if;
        when a_1 =>
        if sum_less_zero_i = '1' then
next_state <= a_0;
        else next_state <= a_2;
        end if;
        when a_2 =>
        if sum_less_zero_i = '1' then
next_state <= a_0;
        else next_state <= a_2;
        end if;
        when others => next_state <= a_0;
end case;
 (ready_o, sum_plus_a_o, sum_minus_b_o, y_o)
<= control;
with state select
control <= "1000" when a_0,
"0100" when a_1,
"0011" when a_2,
"0000" when others;
with state select
states <= "11" when a_0,
"10" when a_1,
"01" when a_2,
"00" when others;
```

Fig. 5 Device's control unit operations

For the synthesis of a bit-stream computer, the Xilinx SPARTAN 3E platform of the XC3S100E series was used, in which approximately 6% of the resources were involved. The maximum frequency is 93.9 MHz. Used 24-bit components SUM1, SUM2 and 16-bit components Count, SUM3.

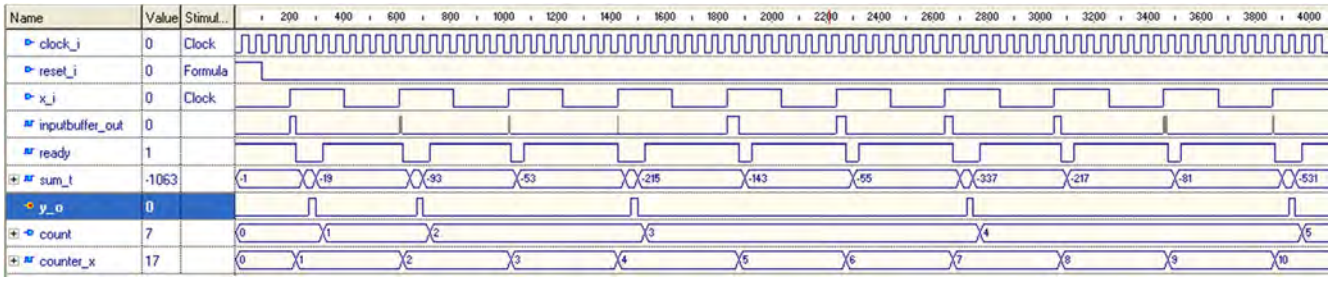


Fig. 6. Results of modeling a behavioral model of a bit-stream power function online computer

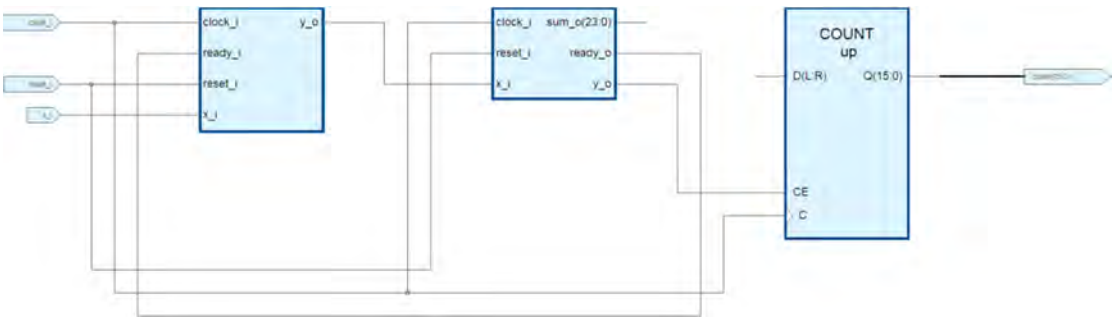


Fig. 7. RTL–scheme of synthesized device

IV. SUMMARY

The article proposes a bit-stream power function online computer with a fractional exponent, in which the operations of exponentiation and root extraction are combined in one device and are performed in real time when bit-stream data arrives at its input.

The scientific novelty lies in the fact that the process of calculating the power function in the device is carried out on the basis of the proposed method for generating increments of increasing step functions, which provides the principle of sampling certain bits from the input bit-stream data, determined by the type of function and the calculation error. This principle allows to increase the frequency of the input bit-stream and provide a minimum computation time. In this case, the absolute error in reproducing the power function is 0.5 units of the least significant bit of the argument.

On the basis of the proposed method, a mathematical model of a bit-stream power function online computer is obtained, which is a system of inequalities. As the main component of the comparison, it is proposed to use an accumulating adder, which compares in parallel codes the increments of two simultaneously reproducible power step functions, the increments of one of which are fed to the adder by means of the bits of the input bit-stream in the direct code, and the increments of the other function - by its output bits in the complement code. The device has a streaming method for online computations with parallel-sequential conversions over the input bit stream.

The hardware implementation is performed by building a model in the hardware description language in a synthesized VHDL subset and subsequent synthesis using the Xilinx CAD tools. The arithmetic unit of the bit-stream computer is implemented on the basis of Moore's finite state machine and is represented by a composition of the operational and control units. A VHDL-model of the device based on the automatic method of describing the device is developed. The results of creating the behavioral model of the device coincided with theoretical calculations. The device model is implemented in the Xilinx Spartan FPGA.

REFERENCES

[1] Dhafer Al-Makhles, Nitish Patel, Akshya Swain. *"Bitstream control system: Stability and experimental application."* Intern. Conf. on Appl. Electronics. Czech Republic, Pilsen, 2013. pp. 1–6.

[2] Bureneva O.I., Zhirnova O.A. *"Bit-potokovoye ustroystvo izvlecheniya kvadratnogo kornya."* [Bit-stream square root extracting device] Izvestiya LETI, №2, 2019, pp. 26 – 32.

[3] Saf'yannikov N.M., Bureneva O.I. *"Sledyashchiy potokovyy vychislitel'nyy preobrazovatel' dlya intellektual'nykh izmeritel'nykh system. "* [Tracking stream computing converter for intelligent measurement systems] Mezhdunarodnaya konferentsiya po myagkim vychisleniyam. 2019, T.1, pp. 263-266.

[4] Peng Li, David J. Lilja, Weikang Qian, Marc D. Riedel, Kia Bazargan *"Logical Computation on Stochastic Bit Streams with Linear Finite-State Machines."* IEEE Transactions on Computers, Vol. 63 , No 6, 2014

[5] A.I. Gulin, N.M. Safyannikov, O.I. Bureneva, A.Yu. Kaydanovich. *"Assurance of Fault-Tolerance in Bit-Stream Computing Converters. "* Proceeding of 16th IEEE East-West Design & Test Symposium (EWDTS'2018). Kazan, Russia, September 14 – 17, 2018. pp. 418 – 421.

[6] M.YU. Stakhiv *Avtoref. dysertatsiyi. "Tsyfrovi funktsional'ni peretvoryuvachi roz·hortuyuchoho typu z pokrashchenymy kharakterystykamy. "* [Abstract of the dissertation. Deployment-type digital functional converters with improved characteristics] Vydavnytstvo Natsional′noho universytetu «L′vivs′ka politekhnika». 2013. 21 p.

[7] Khoa Dang Pham, Edson Horta, Dirk Koch *"A tool and API for FPGA bitstream manipulations."* Design, Automation & Test in Europe Conference & Exhibition (DATE), Lausanne, Switzerland, 2017.

[8] L.V. Larchenko, E.M. Kulak, B.D. Larchenko. *"Funktsional'ne peretvorennya impul'snykh potokiv v aparatnykh obchyslyuvachakh matematychnykh funktsiy."* [Functional conversion of pulse streams in hardware mathematical functions computers], Radioelectronics and Informatics, №3, 2019, pp. 27-34.

Reinforcement Learning for Anti-Ransomware Testing

Alexander Adamov
NioGuard Security Lab /
Design Automation Dep.
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine /
Dep. of Software Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
ada@nioguard.com

Anders Carlsson
Dep. of Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
anders.carlsson@bth.se

***Abstract*—In this paper, we are going to verify the possibility to create a ransomware simulation that will use an arbitrary combination of known tactics and techniques to bypass an anti-malware defense.**

To verify this hypothesis, we conducted an experiment in which an agent was trained with the help of reinforcement learning to run the ransomware simulator in a way that can bypass anti-ransomware solution and encrypt the target files.

The novelty of the proposed method lies in applying reinforcement learning to anti-ransomware testing that may help to identify weaknesses in the anti-ransomware defense and fix them before a real attack happens.

Keywords—ransomware, machine learning, reinforcement learning, artificial intelligence, anti-ransomware testing

I. Introduction

68% of ransomware attacks go unnoticed according to the latest report by US cybersecurity provider FireEye [1] that draws the cybersecurity experts' attention to this problem.

In the previous work, we already analyzed the LockerGoga ransomware used in the targeted attack against Norsk Hydro in consequence of which the company needed to switch to manual operation mode reducing the production capacity. The discovered techniques included digital signing of ransomware executables and multi process encryption when a single worker process was created and responsible for encryption only one user's file. These techniques helped the ransomware go undetected. [2]

The author(s) of Maze ransomware used in the recent attack against Canon complained that it is so easy nowadays to bypass an antivirus protection because they "place a signature on data section in the packer layer" that makes EDR (Endpoint Detection and Response) [3] solutions useless when it comes to detecting targeted ransomware attacks. Because, once repacked, a piece of malware becomes undetectable again.

Another recent ransomware called WastedLocker and operated by the Evil Corp gang has shown the power of bypassing anti-ransomware modules that typically rely detecting anomalous behavior. The ransomware employed Alternate Data Streams (ADS) in NTFS to drop the payload and memory-mapped files for encrypting user's data. As a result, WasteLocker managed to encrypt data stored on the Garmin's servers. [4]

II. Ransomware Attack Simulation

To be able to test if antiviruses (EDR solutions) can detect an unknown ransomware attack, we propose the attack simulation. An example of such approach is MITRE ATT&CK Evaluation project [5]. The first evaluation of EDR solutions was performed based on the discovered attacks by APT29 group attributed to Russian Intelligence Service. The attack simulation included tactics and techniques of this hacking group [6].

Similarly, we designed a ransomware simulation tool to imitate techniques employed by ransomware. For this particular experiment, only three parameters were chosen: 1) adding extension to an encrypted file, e.g. '.enc'; 2) encoding the encrypted data with Base64 that helps to reduce an entropy level; 3) the number of files to be encrypted per step. The simulator uses AES-256 for encryption and targets documents, multimedia files, and archives that are typically encrypted by ransomware. The goal of the Ransomware Simulator is to encrypt the maximum number of files in the minimal number of steps on the target system.

TABLE I
THE CHOSEN PARAMETERS OF THE RANSOMWARE SIMULATOR

Parameter	Value 0	Value 1	Value 2	Value 3
Adding the extension	no	yes		
Base64 encoding	no	yes		
The number of encrypted files per action	1	2	5	10

III. Ransomware Defense Simulation

To counteract the Ransomware Simulator, we created Ransomware Detector. The detector implements three methods to detect the ransomware activity in correspondence with the Ransomware Simulator's parameters: 1) checking if the second extension exists; 2) entropy level evaluation; 3) detection of anomalous modification time of the files (e.g. 8 files have been modified within 1 second).

The detection *Threshold* was set to 8, which means if the Ransomware Detector sees 8 files with one of the following anomalies: 1) second extension; 2) high level of entropy that indicates that data are encrypted; 3) similar modification time then it triggers an alert 'Ransomware Detected' and block the attack. The Ransomware Simulator failed and the game (attack) is over (blocked).

IV. Environment

The recent targeted ransomware attacks such as Maze, WastedLocker, Netwalker, Clop, and others target Windows OS. Therefore, we used a Windows 10 virtual machine [7] as a simulation environment where we placed the folder with ten files that include documents, multimedia files, archives that are typically encrypted by ransomware.

V. Reinforcement Learning

Ransomware Simulator is a tool that requires a set of parameters as an input to operate. The problem is how to find these parameters that will allow the Ransomware Simulator

978-1-7281-1003-5/19/$31.00 ©2020 IEEE

to bypass the Ransomware Detector. To address this problem, we came to the idea of adding Artificial Intelligence (AI) to the Ransomware Simulator with the help of Reinforcement Learning (RL) that should be well known to the players of the Real-Time Strategy (RTS) games, such as StarCraft.

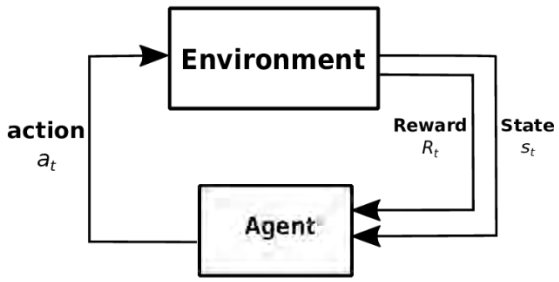


Fig. 1. Reinforcement learning process.

The key advantage of RL is that it does not require training data or specific expertise in the domain. It needs only a goal to be specified and the Agent finds the optimal way (a policy) to achieve that goal using the trial and error method.

In RL, we have the Agent and Environment. The Agent performs actions that affect in some way the Environment and receives the new state of the Environment and the Reward (a numerical score) that evaluates how good the previous action was in terms of leading the Agent to maximization of the total reward in the long run.

In our case the Environment can be a user's Windows OS with antivirus (the Ransomware Detector). The Actions performed by the Agent are ransomware attempts to encrypt user's files. After executing an Action, the Agent receives information about the new state of the Environment and how successful the attempt was. The state of the Environment is evaluated based on the number of encrypted files [8].

Q-learning algorithm is commonly used to solve RL tasks. Q-learning is a reinforcement learning algorithm defined over Finite Markov Decision Process (FMDP) which does not require creating a model of the Environment. The algorithm calculates the quality (Q) of a combination of a state (S) and action (A) based on a reward value (R):

$$Q: S \times A \rightarrow R \qquad (1)$$

The Q values are updated based on the reward the Agent got and the reward the Agent expects next.

$$Q(s_t, a_t) \coloneqq Q(s_t, a_t) + \\ + \alpha \cdot [r_{t+1} + \lambda \cdot max_a Q(s_t, a) - Q(s_t, a_t)] \qquad (2)$$

, where

α – learning rate;

r – Reward;

λ – discount – how the future Action value is weighted over the one at present;

$max_a Q(s_t, a)$ – the estimated Reward from the next Action.

The algorithm requires the definition of the possible states, actions, and reward function.

Set of States. There are 11 states in our model that represents the number of encrypted files in the target folder from 0 to 10.

Set of Actions. There are 16 possible actions that are the combinations of three Ransomware Simulator's parameters.

TABLE I I
THE ACTIONS ENCODING TABLE

Action code	Extension (code)	Base64 (code)	Number of files (code)
0	0	0	0
1	0	0	1
2	0	0	2
3	0	0	3
4	0	1	0
5	0	1	1
6	0	1	2
7	0	1	3
8	1	0	0
9	1	0	1
10	1	0	2
11	1	0	3
12	1	1	0
13	1	1	1
14	1	1	2
15	1	1	3

See Table 1 to translate the codes to the actual values of the parameters.

Learning strategy. The algorithm starts with the random (*exploration*) policy and then slowly reduce the probability of random choice for an action from 1 in the beginning to 0 at the end of the learning process (*exploitation*). Discount = 0.95. Learning rate = 0.1.

VI. RESULTS

After 1000 iterations (595 games have been played), we obtained the following results.

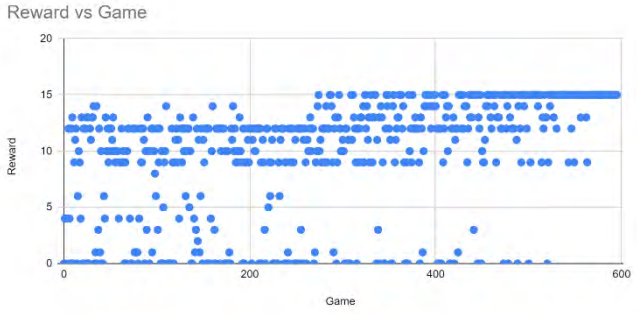


Fig. 2. Learning progress (Reward vs. Game rate).

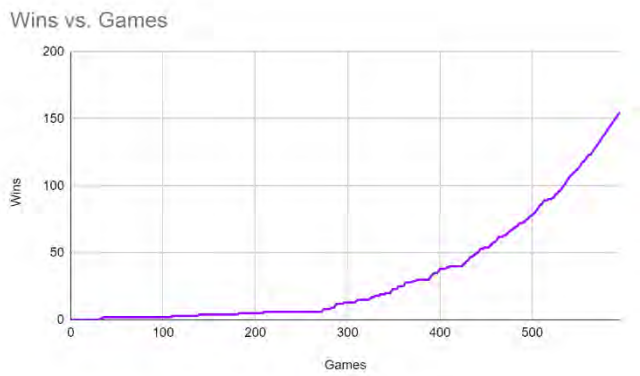


Fig. 3. Learning curve (Wins vs. Games rate).

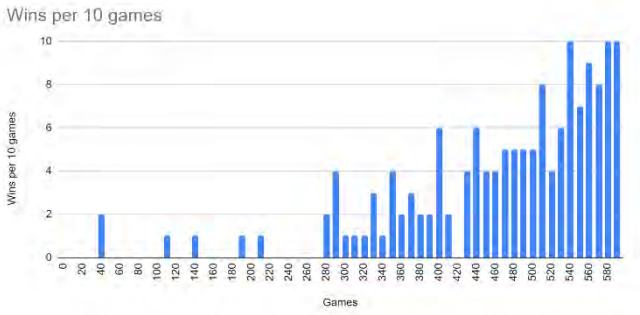


Fig. 4. The number of Wins per 10 games.

Q (S, A)	States								Threshold		
Actions	0	1	2	3	4	5	6	7	8	9	10
0	3.742632	0.787156	0.50037	0	0	1.007811	0.119928	2.686471	1.771443	0	
1	7.105196	1.013578	1.252902	0.625865	0	5.491767	0	0.110781	0.020577	0	
2	7.545032	3.370516	1.118979	0.296561	0	1.170941	0.170201	0.147976	0.215866	0.763711	
3	6.881859	1.036887	1.431128	0	0	1.178761	0.10031	0.526232	0.019609	0.1805	
4	3.948145	0	0	0.231532	0.120758	0.790898	0.599471	0.871498	0.483216	0	
5	5.816442	0	1.158697	0.693802	0	1.778738	0.715084	0.205816	0.029846	0	
6	3.734055	8.923577	2.616297	0	0	0.747448	0.252802	0.236518	0.020577	0	
7	7.022468	0.056858	1.931174	0.563466	0	1.003211	0.213173	0.198259	0.377146	0	
8	4.158711	0.963861	1.045184	0	0	0.459586	0.590549	0.028423	0.28698	0.01805	
9	5.872631	0.463353	0.387103	0	1.524341	1.143632	0.358465	0.771859	0.112953	0.119148	
10	5.959267	0.9285	8.530923	1.739734	0	1.55761	0.141551	0.244138	0.172354	0	
11	6.141237	0	0.768712	0	0.144812	1.640483	0.149433	0.418004	0.055891	0.068169	
12	5.179697	1.992647	0.713766	0	0.443053	0.652553	0.226234	0.173608	0.372956	0.089291	
13	3.181544	1.945088	0.629294	0	0.422005	1.051774	2.283579	0.231025	0.166056	0	
14	14.11996	2.516235	1.889826	3.106087	0.087103	0.705648	0.275448	0.247926	0.20654	0	
15	5.828471	0.885469	0	0.163353	0	1.260362	0.080788	0.186429	0.010403	0.07444	

All the files are encrypted

Fig. 5. Q-table with the optimal policy highlighted.

VII. Analysis of Results

After 1000 iterations (595 games), the Agent found the *optimal policy* that allowed it to encrypt all the ten files in the target folder:

1. *State 0: Action 14* - encrypt 5 files adding the extension and apply Base64 encoding to reduce entropy.
2. *State 5: Action 1* – encrypt 2 files without adding the extension and Base64 encoding.
3. *State 7*: *Action 0* – encrypt 1 file without adding the extension and Base64 encoding.
4. *State 8*: *Action 0* – encrypt 1 file without adding the extension and Base64 encoding.
5. *State 9*: *Action 2* – encrypt 5 files without adding the extension and Base64 encoding.

At State 9, any Action would lead to win because only one file left unencrypted. Moreover, at State 9, we have:

- 5 (or 7 – two consequent iterations may result in the encrypted files having similar modification time within Δt) files modified at the same time (within Δt = 1 sec),
- 5 files with the extension,
- 4 files with high entropy.

None of these exceeds the detection threshold equal to 8.

VIII. Conclusions

The obtained results during the experiments show that RL can help to discover an attack strategy that can overcome a behavior-based anti-ransomware protection. It is worth noting, that the experiment was conducted on the Ransomware Detector that only represents the limited number of detection methods imitating the behavior of a real EDR solution. However, the presented results look promising and the proposed RL approach for anti-ransomware testing can be further applied to the anti-malware products with anti-ransomware modules available on the market.

References

[1] Security effectiveness Report 2020. Deep dive into cyber reality, Mandiant, 2020.

[2] A. Adamov, A. Carlsson and T. Surmacz. An Analysis of LockerGoga Ransomware, 2019 IEEE East-West Design & Test Symposium (EWDTS), Batumi, Georgia, 2019, pp. 1-5, doi: 10.1109/EWDTS.2019.8884472.

[3] A. Chuvakin. Named: Endpoint Threat Detection & Response, Gatner, 2013, available at https://blogs.gartner.com/anton-chuvakin/2013/07/26/named-endpoint-threat-detection-response/

[4] WastedLocker's techniques point to a familiar heritage, Sophos, 2020, available at https://news.sophos.com/en-us/2020/08/04/wastedlocker-techniques-point-to-a-familiar-heritage/

[5] Methodology Overview: Adversary Emulation. MITRE, 2020, available at https://attackevals.mitre.org/adversary-emulation.html

[6] APT29 Emulation. MITRE, 2020, available at https://attackevals.mitre.org/APT29/

[7] Windows virtual machines, 2020, available at https://developer.microsoft.com/en-us/microsoft-edge/tools/vms/

[8] F. Bach, R. S. Sutton, A. G. Barto. Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning series). Second edition, A Bradford Book, 2018.

Phase Shifter Designs Based on Miniature Couplers

Ilya A. Terebov
Ural Federal University
Yekaterinburg, Russia
ilyaterebov@yandex.ru

Denis A. Letavin
Ural Federal University
Yekaterinburg, Russia
d.a.letavin@urfu.ru

***Abstract*—Design solutions for implementing phase shifters based on compact directional couplers and compact phase-shifting cells are considered. Such phase shifters can be used in power supply circuits of antenna arrays at low frequencies. The implementation of the phase shifter on artificial transmission lines has an area on the microwave substrate that is 73.2% less than the implementation on conventional transmission lines. A phase shifter with compact structures occupies 74.5% less space than a conventional design.**

Keywords— coupler, miniaturization, device.

I. Introduction

A phase shifter is a two-port device that implements a phase change from input to output to the desired phase shift value. The signal at the output of the phase shifter will differ not only in the phase shift, but also in the introduced amplitude distortions. The implementation of a phase shifter on a printed circuit Board can be obtained in many ways. In our case, we will consider the construction of a negative type. This design consists of a directional coupler to the outputs of which the sliding sections are connected. When connecting or disconnecting these sections via pin diodes, you can change the phase at the output of the phase shifter. It is also worth noting that the reflected signal from the sections is added in phase at the output, and the input is antiphase summation of signals. To obtain different phase shifts, the phase shifters are combined as a cascade inclusion. Analog or voltage-controlled phase shifters are used in many applications. However, the main use of such devices is found in schemes for the formation of several beams at the antenna array.It is quite difficult to distinguish the classification of phase shifters, due to the large number of classification parameters and device variants. Phase shifters are classified according to the type of control, the principle of operation, the method of inclusion in the path, as well as the nature of the phase change. The area occupied by the phase shifters on the sub-spoon will be the larger the lower the frequency. The IEEE has articles that describe information about compact coupler designs that can be used in reflective phase shifters. The appearance of compact couplers is very diverse. Various implementation com-Pact couplers are described in [1]-[12]. In [1], the authors propose a new design of a compact coupler obtained by replacing homogeneous transmission lines with their analog in the form of inhomogeneous lines. In work [2], a method of miniaturization is presented, which consists in shortening the transmission line due to slowing down of the wave by discontinuous structures. Work [3] describes a compact coupler obtained by using artificial transmission lines. The work [4,5] presents a compact circular directional coupler operating at two frequencies due to the use of artificial transmission lines and additionally installed transmission line segments. A directional coupler with high miniaturization rates is presented in [6], in which the authors use P-circuits with the combination of extreme capacitances. In [7], a method of miniaturization is proposed, which consists in using quasi-lumped elements and T-shaped circuits. Work [8] describes a method of miniaturization based on fractal lines. In [9], the area of the coupler is reduced by the usual bends of the transmission lines. Work [10] describes the efficiency of using two coupled segments instead of a conventional line to miniaturize a coupler. In [11], the possibility of reducing the area of the device by using low-pass filters was demonstrated.Depending on the purpose and frequency range, their parameters and phase shifters can differ very much. This paper presents two implementation options for compact phase shifters that can be used in power supply circuits for antenna arrays.

II. Design Shifter

The target of this work is to develop models of compact phase shifters in the Cadence AWR program and then test the calculations in practice using high-precision equipment. To achieve this goal, the AWR initially modeled the phase shifter as standard. The layout of such a phase shifter is illustrated in Fig.1. It consists of a standard coupler and two phase-shifting sections made on conventional transmission line segments.

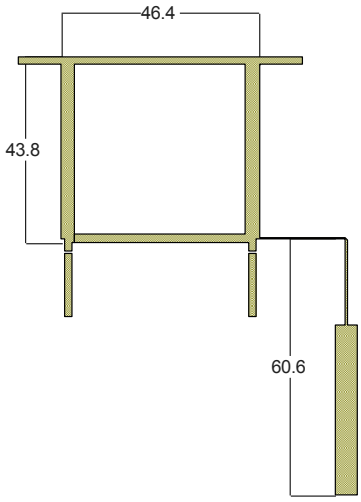


Fig. 1. The topology of a conventional phase shifter obtained in AWR

The calculated graph s-parameters from frequency is presented in Fig.2. The area of the device will be measured by area, without directional coupler output lines. With this calculation, the size of the phase shifter at a frequency of 1 GHz is equal to 2032 mm^2.

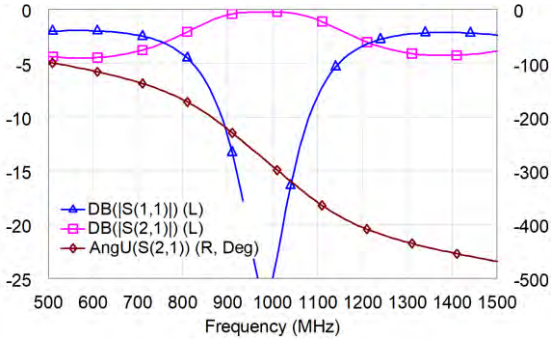


Fig. 2. Frequency characteristics of a standard phase shifter

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

Based on the obtained topology of the phase shifter in AWR, it can be seen that the device has a large area, and the area inside the coupler is not used in any way. Therefore, it is proposed to replace the segments with its analog in the form of a T-section consisting of LC elements. This section is a low-pass filter and when selecting the necessary parameters, it can provide equivalent characteristics at the Central frequency and its vicinity with the characteristics of the segments to be replaced. Initially, all quarter-wave segments were replaced with microstrip cells, implemented in the form of high-resistance lines and low-resistance sections connected to them. The implementation of a phase shifter based on such a circuit design is shown in Fig.3. In Fig.4 shows the design characteristics of such a phase shifter without connecting the phase-shifting sections. The area of the device is 543 mm^2.

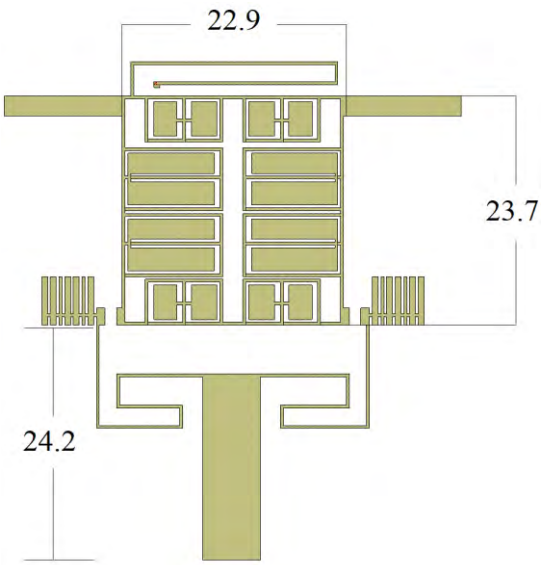


Fig. 3. Phase shifter based on microstrip cells

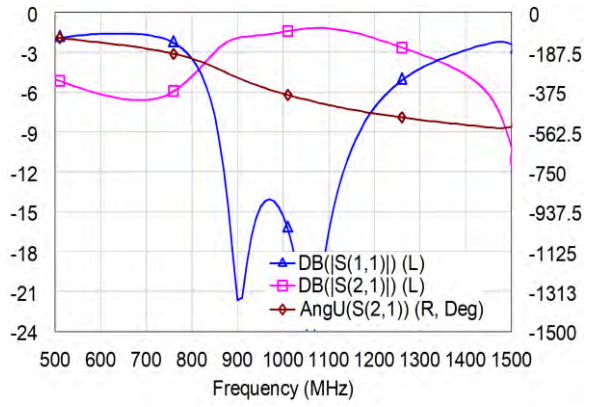


Fig. 4. The frequency characteristics of the miniature phase shifter on the basis of cell

Additional reduction of the area is possible by replacing the power filter with concentrated chip elements, as well as using a short circuit at the ends of the fan-moving cells, which will remove the loop for grounding. It can be seen that the replacement of quarter-wave segments led to an increase in losses at the phase shifter and a deterioration of the alignment, which is associated with a denser packing of elements with each other, which resulted in parasitic connections. It is also worth noting that microstrip cells are low-pass filters, which have a transfer coefficient that differs from the transmission coefficient of conventional sections. However, the use of microstrip cells allowed to reduce the area of the device 3.75 times.

The amount of phase change at the output of the phase shifter will depend on the electrical length of the connected sections. The next implementation option for a miniature phase shifter is to replace standard segments with compact structures made using microstrip technology. Such structures are also implemented on LC elements, and the topology of the resulting phase shifter based on them is presented in Fig.5. the size of the device is 517 mm2, which is 3.9 times less than the standard implementation of the phase rotator.

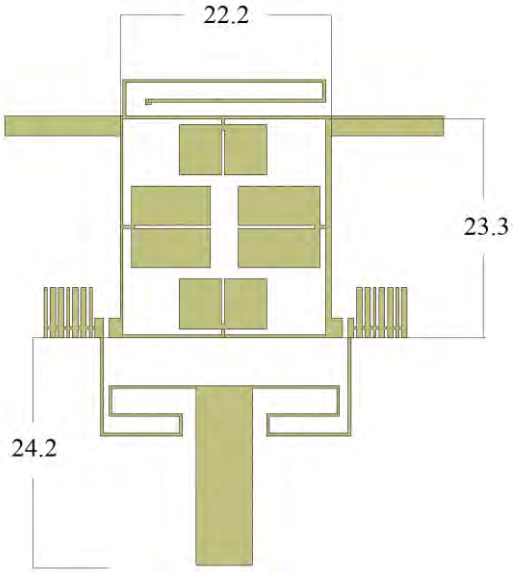


Fig. 5. Phase shifter based on compact structures

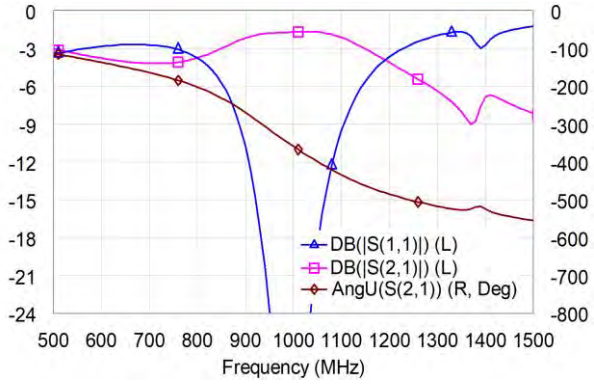


Fig. 6. Frequency characteristics of a miniature phase shifter based on structures

The obtained characteristics show that the use of compact structures in comparison with microstrip cells is more profitable, since there are fewer parasitic connections between elements, it is easier to configure, and there are no gaps between neighboring elements in one structure. However, the use of microstrip cells allowed to reduce the area of the device 3.75 times. The amount of phase change at the output of the phase shifter will depend on the electrical length of the connected sections.

III. Prototype Measurements

The developed models of miniature phase shifters were manufactured and measured using the rode vector circuit

analyzer&Shwarz ZVA24. Figure 7.9 shows a cell-based phase shifter layout. The measured characteristics are shown in figures 8.10.

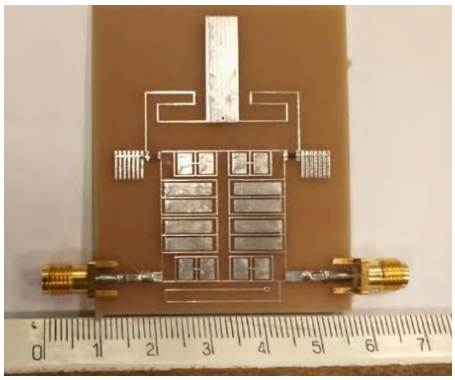


Fig. 7. The prototype of a compact coupler

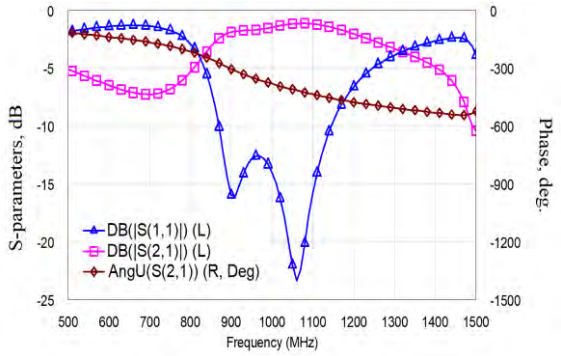


Fig. 8. Measured characteristics of a miniature coupler based on microstrip cells

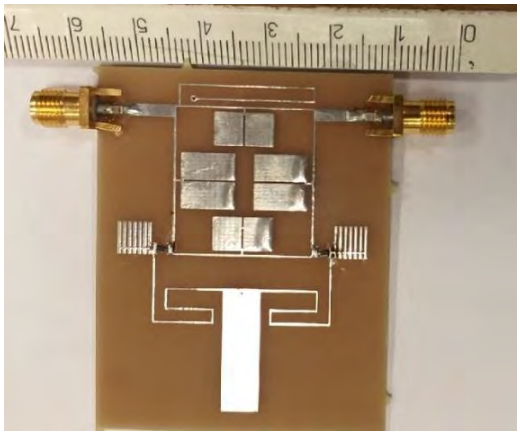


Fig. 7. The prototype of a compact coupler

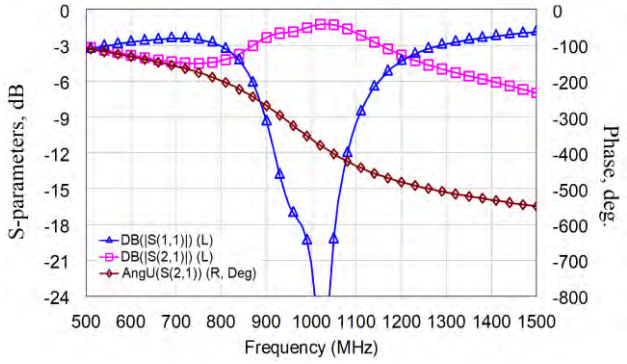


Fig. 8. Measured characteristics of a miniature coupler based on structures

The obtained characteristics showed good convergence with theoretical calculations, which indicates the correctness of designing and obtaining characteristics of compact devices at the level of conventional design.

IV. Conclusion

In this work, the phase shifter assembled in the traditional AWR program was miniaturized. Two models of miniature phase shifters obtained by replacing segments with microstrip cells and compact structures are shown and studied. Cells and structures play the role of a low-pass filter, and have similar characteristics to the transmission line segments in the phase shifter's bandwidth. Without taking into account the output lines, the area of the phase shifter based on cells was reduced by 3.75 times, and based on structures by 3.9 times. Phase change in miniature phase shifters is implemented in the same way as in a conventional design, by connecting or disconnecting the phase-shifting sections through the controlled elements. The only difference is that these sections are made in the form of a slowing structure.

Acknowledgment

The work was supported by Act 211 Government of the Russian Federation, contract № 02.A03.21.0006.

References

[1] F. Hosseini, M. Khalaj-Amir Hosseini and M. Yazdani, “Novel compact branch-line coupler using non-uniform transmission line,” 2009 Asia Pacific Microwave Conference. DOI: 10.1109/APMC.2009.5384390

[2] Mehdi Nosrati, and Salman Karbasi Valashani,” A Novel Compact Branch-Line Coupler using four coupled transmission lines” , progress in Microwave and Optical Techmology Letters / Vol. 50, No. 6, June 2008.

[3] C.-L. Hsu, J.-T. Kuo, and C.-W. Chang, “Miniaturized dualband hybrid couplers with arbitrary power division ratios,” IEEE Trans. Microw. Theory Tech., vol. 57, no. 1, pp. 149-156, Jan. 2009.

[4] Chi-Feng Chen, Sheng-Fa Chang and Bo-Hao Tseng, “Compact Microstrip Dual-Band Quadrature Coupler Based on Coupled-Resonator Technique,” IEEE Microwave and Wireless Components Letters. DOI: 10.1109/LMWC.2016.2575006

[5] K.-S. Chin, K.-M. Lin, Y.-H. Wei, T.-H. Tseng, and Y.-J. Yang, “Compact dual-band branch-line and rat-race couplers with stepped-impedance-stub lines,” IEEE Trans. Microw. Theory Tech., vol. 58, no. 5, pp. 1213–1221, May 2010.

[6] Chi-Hsing Wu and Chao-Hsiung Tseng, “A compact branch-line coupler using π-equivalent shunt-stub-based artificial transmission lines,” 2010 Asia-Pacific Microwave Conference.

[7] S.-S. Liao and J.-T. Peng, “Compact planar microstrip branchline couplers using the quasi-lumped elements approach with nonsymmetrical and symmetrical T-shaped structure,” IEEE Trans. Microw. Theory Tech., vol. 54, pp. 3508–3514, Sep. 2006.

[8] H. Ghali, and T. A. Moselhy, “Miniaturized fractal rat-race, branch-line, and coupled-line hybrids,” IEEE Trans. Microw. Theory Tech., vol. 52, pp. 2513–2520, Nov. 2004.

[9] Ashmi Chakraborty Das, Lakhindar Murmu and Santanu Dwari, “A compact branch-line coupler using folded microstrip lines.” 2013 International Conference on Microwave and Photonics (ICMAP). DOI: 10.1109/ICMAP.2013.6733485

[10] V. Velidi, G. Shankar, K. Divyabramham, and S. Sanyal, “Compact coupled line quadrature hybrid coupler with enhanced balance bandwidth,” Applied Electromagnetics Conference, pp. 1 – 4, 2011.

[11] Letavin, D.A., Mitelman, Y.E. and Chechetkin, V.A., “A Novel Simple Miniaturization Technique for Double Rat-race coupler,” 2019 Antennas Design and Measurement International Conference, ADMInC 2019. DOI: 10.1109/ADMInC47948.2019.8969084.

Monitoring and Control System of Three-Phase Electrical Loads on Railway Trains

Sergei A. Kalabanov
Department of Radio Physics
Institute of Physics
Kazan Federal University
420008, 18th Kremlyovskaya Str.,
Kazan, Russian Federation
kazansergei@mail.ru

Rinat I. Shagiev
Department of Radio Physics
Institute of Physics
Kazan Federal University
420008, 18th Kremlyovskaya Str.,
Kazan, Russian Federation
r3ntil@gmail.com

Rashid A. Ishmuratov
Department of Computer Science,
Institute of Digital Technology and Economics
Kazan State Power Engineering University
420066, 51st Krasnoselskaya Str.
Kazan, Russian Federation
rash-i@mail.ru

Michael V. Onischuk
"Dalreftrans" Ltd,
3d Mytishchinskaya Str., 16
Moscow, Russian Federation
MOnischuk@fesco.com

***Abstract*—The article describes an automated system for remote monitoring and control of the power supply of refrigerator containers, which are installed on moving railway coupled platforms. Power Line Communication technology is implemented for data transferring in the system. The principles of construction of the automated system are considered, its remote peripheral units and the central control unit are described. A structural-functional diagram of the electrical control system, the main technical characteristics and features of the hardware and software implementation of the developed devices are presented.**

Keywords—Electronics, Electrical Loads, Power Line Communication, Software, Sensors, Data collection, Railway Trains, Refrigerator Container

I. INTRODUCTION

The operation of refrigerator containers requires constant monitoring of three-phase power supply availability. In the case of the power supply phase's damage, the self-protection circuity will not allow restarting the refrigerator again in order to avoid possible failures of the expensive equipment. Additional failures may appear during the operation of refrigerator containers on railway platforms as part of coupled wagons. In this case, all containers receive three-phase power from one autonomous diesel generator of limited power. The generator will inevitably experience overloads at the moment of its launch if all refrigerators are in "on". The implementation of automatic control system for power supply phase's failure detection and for remote on/off the refrigerators will solve such problems and ensure the operation of expensive refrigeration equipment in normal mode, which in turn extends their service life. In addition, the automated system allows real-time signaling to the operator of possible emergency situations or critical operating modes (open circuit or contact's failure).

The ready-made data collection and control systems offered in the modern electronic industry market do not allow taking into account all the features and needs of a particular production or enterprise, or do not provide the implementation of the necessary additional functions [1, 2]. Thus, the company [1] specializes in data acquisition and control systems using PLC technology. They allow collecting various data from refrigerated containers, such as temperature, humidity, air flow rate, etc. However, these systems do not allow real-time monitoring of the state of the power phases and promptly and remotely turn off the power of refrigerated containers in the event of a possible loss of phase's power. In addition, the collected data is transmitted to the special control center. At the same time, in order to use this data, the user must, in addition to the purchased equipment, pay for the corresponding electronic subscription, which may turn out to be economically unprofitable for a potential customer.

Thus, in the presence of a wide industrial market of ready-made data acquisition and control systems, the importance for the developing of new original systems adapted to a specific production and considering all the customer's wishes remains. It is also important to technically support the system, the possibility to upgrade it as it operates in real production conditions, as well as the attractiveness of the price of the finished system for the customer.

The purpose of this work is to develop the new original automated system for monitoring and control of three-phase electrical loads on moving railway platforms. In addition to collecting data, the system should provide the ability to continuously monitor the state of the power phases and quickly and remotely turn off the power of refrigerated containers in the event of a possible power phase failure. For this, it is necessary to solve the following tasks:

1) to develop the general structure of the information-measuring and control system (hereinafter, the system) and to determine the optimal technology (method) of data transmission in the system;

2) to develop hardware and software parts of remote (peripheral) executive and central units of the system;

3) to develop a remote controller with a special keyboard and LED indicators for remote power on/off for each refrigerator container on the platform as part of coupled wagons;

4) to develop software for the central unit of the system for exchanging data with peripheral units and for providing supervisory load control functions.

II. Choice of Data Transfer Technology and a General Description of the System

In this work, based on our previous design experience of similar data transmission systems [3, 4, 5], the choice was made in favor of the PLC technology – the transfer of digital data over power supply lines [6, 7]. This technology, developed several decades ago, continues to attract the attention of many industrial companies [8]. When choosing a data transmission technology, we took into account the such factors as the spatial distance of the railway platforms from the diesel generator and the overall level of electromagnetic interference in the power line of the moving railway train were taken into account. The validity of our choice was confirmed later in the course of the preliminary testing of the developed equipment on the railway coupled wagons, as well as the experience of actual operation of the train on the railway link Moscow-Vladivostok.

Railway coupled wagons includes up to 20 cargo platforms, where 1 or 2 refrigerator containers are installed, and one power platform (in the middle of the train), where a diesel generator and a separate room for service personnel (operators) are installed.

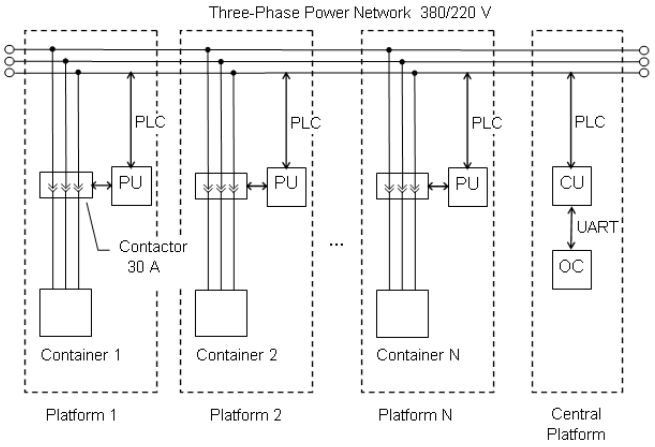


Fig. 1. Structural-functional scheme of the data collection network

A common three-phase power supply cable 380/220 V runs along the all platforms from the autonomous diesel generator. For each platform from the backbone cable there are branches for power supplying to each refrigerator container and the electrical equipment (socket, fuse and contactor of 30A for manually turning on/off the power).

The automated system consists of remote peripheral units (PU) designed to be installed on each railway platform, one central unit (CU) and an operator's console (OC) connected to CU to control loads in the system (Fig. 1).

Each PU is installed on the platform near the contactor supplying power to the refrigerator container and it is connected to the mains. PU has three voltage sensors of each phase of the three-phase power supply and a low power relay, which controls the power supply of the refrigerator container via the contactor.

The central unit is located in the room of operators. Its function is to collect information from remote PUs and coordinate the work of the entire system. The function of the CU includes also the transfer of data to the operator's console, which has a set of special buttons and LED indicators.

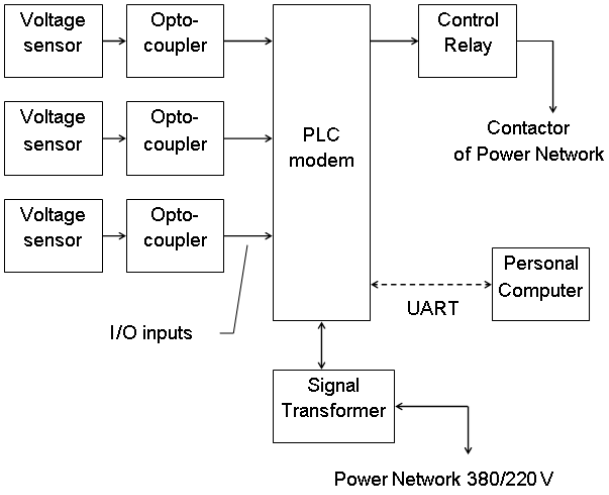


Fig. 2. Structural-functional scheme of the remote peripheral unit of the system

This panel is the main tool with which the operator carries out operational and complete control and analysis of the work of all refrigerator containers in the train.

Data transmission in the system is carried out through a backbone line of 380/220V using PLC technology. For more reliable PLC signal transmission through the power cable, a special phase-to-phase mixer is additionally connected to the individual phases of the power network, which includes two 0.47 μF capacitors (the nominal value is selected for the best frequency transmission conditions of the PLC signal 95-120 kHz).

The peripheral unit (PU) was developed on the basis of the IT 700 PLC modem of Yitran Communication [9]. The modem is based on the 8051 microcontroller with the corresponding command system. The structural-functional scheme of PU is shown in Fig. 2. The appearance of the PU (front and back panel) is shown in Fig. 3 Three voltage sensors (for each phase of the power network) are connected to the PLC-modem. The signals from the sensors go to the digital inputs of the PLC-modem through optical isolators in order to galvanically isolate the power mains from the electronic circuit of the modem. The PU includes an electromechanical relay that gives the power supply to the coil of a three-phase contactor. This is how works remote on/off of the power loads (refrigerator containers).

The modem's electronic circuits are connected to the power mains via a matching signal transformer. Control and data exchange in the information system is carried out according to the PLC-modem protocol [9]. In each PU is implemented a queue for sending information packets and checking for successful packet delivery, which allows to increase the reliability of communication between the central and peripheral units.

Initial setup and debugging of the PLC-modem during installation and testing of the system is carried out using a laptop, which is connected via COM port to the UART interface.

The central unit is implemented on the Atmel SAM3S processor [10] in conjunction with a PLC modem, which in this case operates as the network coordinator (Fig. 4-5). The CU performs data collection and general control of

information network. The PLC modem is directly connected to the SAM3S processor via the UART port.

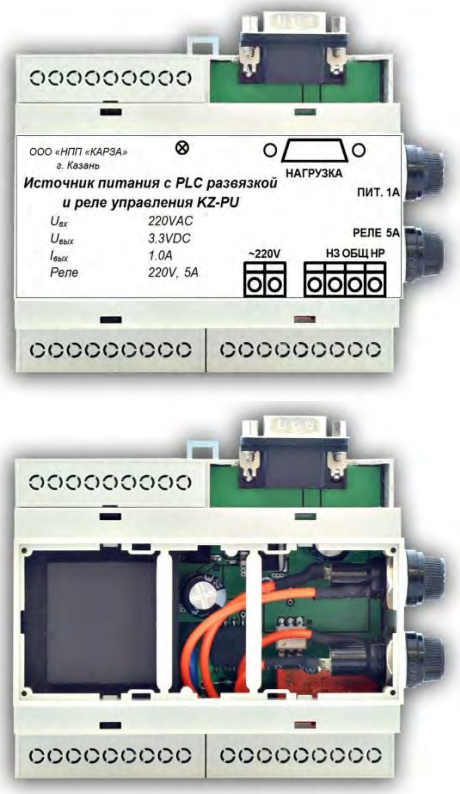


Fig. 3. Appearance of the remote peripheral unit of the system

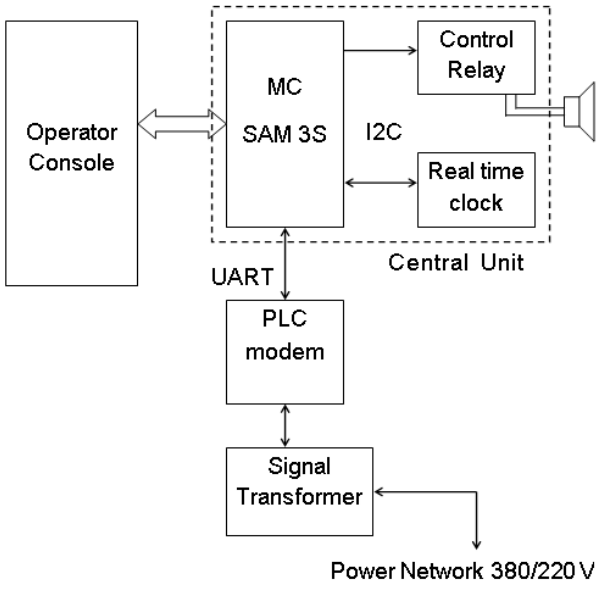


Fig. 4. Structural-functional scheme of the central unit (coordinator of the data collection network)

The software of the CU includes several levels - network and application levels. At the network level, the real-time operating system FreeRTOS [11] was used to ensure interaction with a large number of devices (sensors, microcircuits and relays). This allowed to realize all necessary software processes (information flows) for the data network and to organize their interaction.

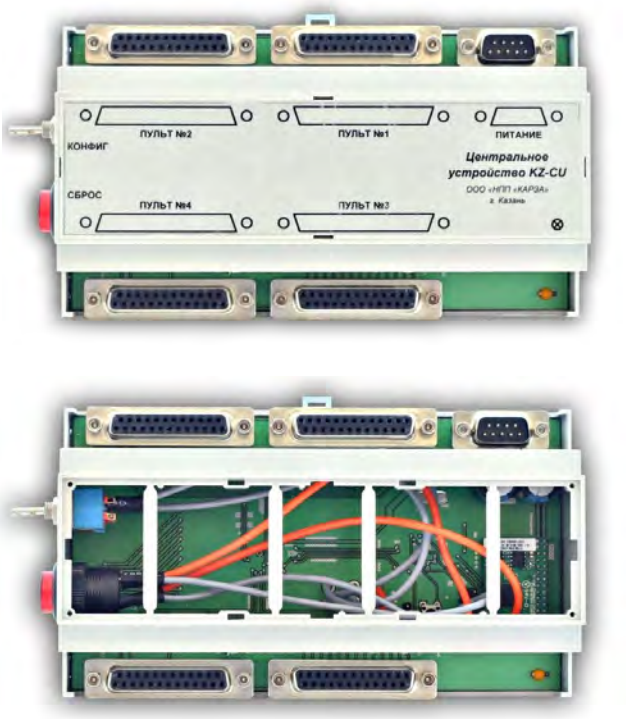


Fig. 5. Appearance of the central unit

At the application level the CPU runs under the control of a specially developed program for the SAM3S processor. The program performs the initial configuration of the modem (coordinator mode, network number, network capacity, etc.), as well as complete control of the entire system (reading the collected data from remote units, etc.). The tuning of the CU and the PLC-modem is carried out using a PC, which is connected via COM port to the UART interface.

A special operator's console (OC) is connected to the central unit to control the loads and to monitor the current state of the working equipment of the moving railway train (Fig. 6). The mechanical buttons are processed by a specialized keyboard processor TCA841 from Texas Instruments. The console also provides a signaling function using a set of LED indicators and they receive the corresponding information signals through the LED drivers STP16CP05TTR from STMicroelectronics [12].

III. Implementation and Experience of Operating the System

After installation of the system equipment on the railway platforms and tuning individual nodes and a data transmission network, the automated system is ready for operation. From this time, the main tool to control the system is the operator's console.

The console includes a set of 80 buttons - 2 buttons ("on" and "off") for each of the possible 40 containers. Each pair of buttons transmits a command to a remote PU for switching the power circuit (see Chapter 2). The console also provides a signaling function using a set of 80 LEDs – 2 for each container. One LED indicates the switching state of the load (on / off), and the other is used to signal a possible emergency situation on the container (phase loss or contact failure). Emergency warning lights are also duplicated by audible alarm (Fig. 4).

The monitoring and control system of three-phase electrical loads for moving railway coupled wagons passed successful tests and was installed for continuous operation on 4 railway trains.

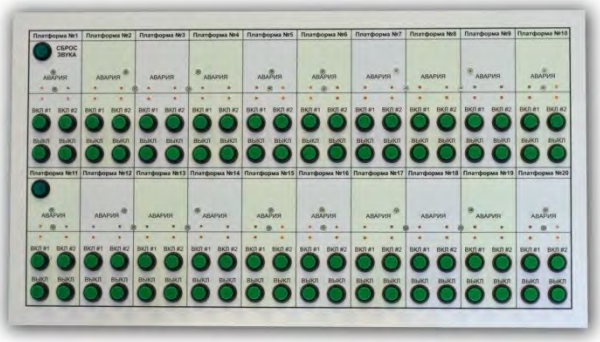


Fig. 6. Appearance of the operator's console

The trains belong to the transport company “Dalreftrans” Ltd (Moscow branch), which specializes in long-distance railway transportation of perishable goods in refrigerator containers [13]. Each railway train contains 20 platforms (10 platforms on each side of the diesel generator) and 1 wagon with a diesel generator and a room for service personnel. Each platform has 2 refrigerator containers. The system for railway train is designed for operation on long railway link (Moscow – Vladivostok, Kaliningrad – Vladivostok).

IV. Conclusion

The monitoring and control system of three-phase electrical power loads (refrigerator containers) for moving railway coupled wagons was developed and implemented based on the technology of digital data transmission over power supply lines - Power Line Communication (PLC). The hardware and software of the peripheral and central units of the system has been developed. The list of functions of the system includes not only monitoring and controlling on/off remote refrigerator containers, but also operational monitoring and alarming of possible emergency conditions of each refrigerator container. Thus, the system allows displaying in real-time mode the current situation of refrigeration equipment on the moving railway coupled wagons for rapid response or analysis for the operator.

The operating experience of the system on the Moscow-Vladivostok railway link, organized by the transport company “Dalreftrans” Ltd, has demonstrated its reliability and efficiency in the context of the railway train that is constantly moving over long distances. Thanks to the individual power management of the refrigerated containers, the system can reduce diesel consumption for the generator by up to 30%. In addition, deterioration of the generator during start-up is prevented, which accordingly extends its service life by about two times (rough estimate).

The developed system can be easily adapted and upgraded to solve any other tasks that require collecting data from remote sensors and devices to a central point for monitoring and control, including the ability to remotely access the system via the Internet. Detailed information about the designer company of the system is given in [14].

Acknowledgment

This work was funded by the Russian Government Program of Competitive Growth of Kazan Federal University.

References

[1] [Electronic resource] Official website of the company “Industrial-Strength Powerline Communications” – Access mode: http://www.adaptivenetworks.com

[2] [Electronic resource] Modern methods of production control // CADmaster. Information-analytical electronic journal. № 3 (43). 2008 – Access mode: http: //www.cadmaster.ru/magazin/articles/cm_43_foreman.html

[3] Kalabanov S., Shagiev R. Ishmuratov R. Automated Data Acquisition System from Industrial Machines // Proceeding of 2018 IEEE East-West Design & Test Symposium (EWDT), Kazan, Russia. - 2018, IEEE Xplore. Digital Library, pp.1-5. – Access mode: https://doi.org/10.1109/EWDTS.2018.8524689

[4] R.I. Shagiev, A.V. Karpov, S.A. Kalabanov, and R.R. Fatykhov, “Data management and data acquisition system based on GSM channel”, Polzunov herald, Barnaul, Russia, 2013, No.2, pp. 214-218. (in Russian)

[5] R.I Shagiev, A.V. Karpov, S.A. Kalabanov, R.A. Ishmuratov, and R.S. Syraev, “Energy-saving automatic heating system for car engines in winter conditions”, Power engineering of Tatarstan, Kazan, Russia, 2015, no.2, pp. 66-72. (in Russian)

[6] Vivek Akarte, Nitin Punse, Ankush Dhanorkar, “Power Line Communication Systems”, International journal of innovative research in electrical, electronics, instrumentation and control engineering, Vol.2, Iss.1, 2014. – Access mode: http://https://www.researchgate.net/publication/263239587_Power_Line_Communication_Systems

[7] K. Verkhulevskiy, “Transmission of information on power supply networks with the help of the Semtech company chip”, Components and technologies, Russia, 2015, no. 11, pp. 50-54. (in Russian)

[8] [Electronic resource] Official website of the group of companies promoting Power Line Communication technology (HD-PLC Alliance) – Access mode: http://www.hd-plc.org/en/

[9] [Electronic resource] Microcontroller with Yitran PLC-modem – Access mode: http://www.yitran.com/ index.aspx?id=3395

[10] [Electronic resource] Official website of the company ATMEL – Access mode: http://www.atmel.com

[11] [Electronic resource] Operating system FreeRTOS Reference Manual – API Functions and Configuration Options – Access mode: http://www.freertos.org

[12] [Electronic resource] Official website of the company STMicroelectronics – Access mode: https://www.st.com/content/st_com/en.html

[13] [Electronic resource] Official website of the company “DalRefTrans” Ltd. – Access mode: https://www.dalreftrans.ru/

[14] [Electronic resource] Official website of the company “Karza” Ltd. – Access mode: http://karza.ru/

The Study of Dynamic Parameters of Corporate Graphic Stations Using Methods of Adaptive Regression Multi-Parameter Modeling

Alexey Andreev
Institute of Physics
Kazan Federal University
Kazan, Russia
alexey-andreev93@mail.ru

Yury Nefedyev
Institute of Physics
Kazan Federal University
Kazan, Russia
star1955@yandex.ru

Natalya Demina
Institute of Physics
Kazan Federal University
Kazan, Russia
vnu_357@mail.ru

***Abstract*—Currently, when solving a number of geophysical and cartographic tasks, one uses corporate graphic stations (CGS) that have particular software packages and digital databases. CGS are used due to the presence of licensed software and authors developments whose installation on several personal computers is not economically and strategically viable. At the same time, CGS may represent a limited access server. Obviously, widening of CGS functions leads to the rise in the number of users. Correspondingly, the increase in the number of CGS users leads to the worsening of software resources usage. To optimize the work, it is necessary to investigate the traffic dynamics (TD) for CGS. The traffic dynamics analysis may be performed using robust methods. For this purpose, one constructs mathematical models of TD for CGS. The aim of this paper is to analyze TD for CGS using the adaptive regression modeling and to find efficient prediction parameters for CGS work. To solve this task, we used adaptive regression multi-parameter (ARMP) modeling. Within ARMP approach, several multi-parameter iterations for assessing the data on time series (DTS) of CGS activity are performed. During the iterations, one finds the most efficient structure DTS, determines the efficiency of adapting the observed values to model ones (ε), and assesses the prediction parameters ($\Delta\varepsilon$). At harmonic analysis of DTS, 2 main harmonics with periods of 1 day and 6 months were selected. At 1-day period, CGS workload gradient starts increasing at 8 a.m. and achieves maximum at noon decreasing by 10 p.m. The study of the main and other harmonic terms when analyzing DTS will allow increasing the efficiency of using CGS and developing a progressive system of TD.**

***Index Terms*—adaptive regression modeling, multiple analysis, graphic stations for cartography and geophysics**

I. Introduction

Currently, cyber-physical systems (CPS) and technologies are successfully used while implementing space missions and creating coordinate and time reference systems. To solve the problems of space orientation and to apply on-board sight cameras and goniometers for this purpose, it is necessary to implement referencing to the visible limb of a celestial body and determine dynamic parameters from long-term series of observations containing large data array and also erroneous measurements. In this process, the use of robust methods for assessing produced values of the desired parameters plays an important role. This paper suggests a noise-immune Hubers method (Huber M estimator method – HMEM) for estimating parameters. The time series are therefore described by complex system of conditional equations of desired parameters whose solution by the classic least squares method cannot eliminate erroneous observations from the processing. It is more plausible to estimate long-term observational series using HMEM. As a result, the values of selenophysical characteristics are obtained with a high accuracy of their estimation.

Mathematical models of time series in technical applications (models of technogenic time series – TTS) play an important role, when solving the tasks of forecasting and increasing the management efficiency. Currently, for TTS simulation one uses: methods of describing a system with polynomials and other regression models; harmonic analysis; method of group consideration of arguments, etc. Nevertheless, in the modern approaches to TTS processing, the mathematical apparatus available in theory for the description of TTS at heterogeneous non-stationarity is not used, the identification and diagnostic test of various regularities underlying the structure of the time series are not performed either. As a result, this leads to decrease in the accuracy of mathematical description and forecasting the dynamics of TTS.

Currently, there is no universal software package for TTS processing that would allow: 1) solving the problems of determining forecast parameters of the system dynamics; 2) increasing the accuracy of describing the model; 3) having a system for time series processing automatization, when constructing multicomponent models of TS.

In this connection, the use of ARMP approach for TTS analyzing, proposed by S. Valeev and practically approved in geophysical applications [1], is relevant. When using the ARMP approach, TTS are described by multicomponent (complex) mathematical models. The application of ARMP approach allows for a significant increase in the accuracy and efficiency of data processing. The ARMP approach represents an implementation of multistep structural and parametric identification. At each step of its application, the construction and analysis of the corresponding component of TTS are performed, its approximation and forecasting accuracies are estimated,

properties of residue are diagnosed, and their adaptation is performed if necessary. To apply the ARMP approach, a base of functions – a set of competing mathematical structures – is required. In the present paper, the analysis of simulating and forecasting for a number of web traffic time series using the ARMP approach, during which estimates parameters of the trend component with the robust Huber method are found, is conducted.

II. The Huber Method as the ARMP Approach

The method based on the maximum likelihood is considered to be the most efficient. For the first time this method was proposed by Huber [2]. The estimates of this method he named M-estimators. The method was developed in a number of subsequent work [3]. The M-estimators method is relatively simple in comparison with the other methods. Under certain conditions (such as absence of wrong measurements, independent measurements, etc.) the results obtained by this method and by LSM coincide. The M-estimators are widely used in data processing that contains gross blunders. In fact, when using this method, all of the measurements obtained in the experiment are processed, and at the analysis each of the measurements has a "weight" value based on the chosen Huber ψ-function.

Estimation parameters using the least square method presumes that the measurement error of model is described by the normal law with predetermined mathematical expectation $E(\varepsilon)$ and by covariance matrix D known with an accuracy up to the certain positive factor σ^2.

$$D = \sigma^2 R, \tag{1}$$

where R is predetermined positive definite matrix. If $E(\varepsilon)$ is predetermined, then the estimation problem using LSM can be reduced to the form with $E(\varepsilon) = 0$ [3]. Thus, LSM problem can be solved assuming that

$$E(\varepsilon) = N[0, \sigma^2 R], \tag{2}$$

$N[0, \sigma^2 R]$ is known expression for normal distribution density.

It is well known that the classic least square method formula has the following form [4]:

$$\Delta q' = S A^T D^{-1} Z, \tag{3}$$

where $S = (A^T D^{-1} A)^{-1}$ is matrix with diagonally situated variance estimations for the vector's $\Delta q'$ components.

A significant disadvantage of least square method is excessive sensitivity of estimations to uncontrolled deviation from the distribution law of the measurement errors [5]. The importance of a deviation from the normal state just as the abnormal measurement errors was noted by Newcomb at the end of the last century. As the usual LSM does not take into account the appearance of emissions possibility, practically empirical and semi-empirical methods of preliminary information cleaning to eliminate the errors are used. Technogenic time series are not exceptions in this regard. The revision of observational material has always been important during a reduction. However, working with a large number of information includes both wrong eliminations and wrong saves.

To remove the influence of the above-mentioned errors, one can use robust version of variation-weighted LSM in which the weight matrix P can be found through Hubers ψ-function. The solution can be found using the following formula [3]:

$$\Delta q = (A^T P A)^{-1} A^T P Z, \tag{4}$$

where

$$diag\ P = \psi(\xi)/\xi, \tag{5}$$

$$\psi(\xi) = \begin{cases} \xi, |\xi| \le b, \\ b\ sign(\xi), |\xi| \ge b, \end{cases} \tag{6}$$

$$\xi = (Z_i - A_i \Delta q_0)/M. \tag{7}$$

In its turn, the median among non-zero value $|Z_i - A_i \Delta q_0|/0.6745$, $M = med\{|Z_i - A_i \Delta q_0|/0.6745 \neq 0\}$.

Δq_0 is preliminary estimate of the Δq vector, b is setting parameter. It should be noted that, as an estimation of the scale, we use the median absolute deviation, which is why for normal distribution consistency we divided the expression by $0,6745$ [2]. The accuracy estimation of the final result Δq is described by the covariance matrix:

$$Cov = k(A^T A)^{-1}, \tag{8}$$

where

$$k = (Mn)^2 \sum_1^n \psi^2(\xi)/(n-m) \left[\sum_1^n \psi^2(\xi) \right]^2. \tag{9}$$

In the expressions (8) and (9) Δq, the final vector value, was used as Δq_0.

III. Subject of Research: Models of the Server Node Traffic Dynamics

In the paper, the data on traffic during the period from April 1, 2016 to April 30, 2017 (9476 observations) is investigated.

Fig. 1a shows a dependence of the original traffic series on time. Fig. 1b presents a curve of the series approximation combined with the model at the same period; the steps of its obtaining are described below.

The time series studied is non-stationary. Hölder exponents $h(q) > 0.8$ are obtained by multifractal analysis; consequently, the dynamics of series correlate, and TS has a regular component.

According to the curve in Fig. 1a, peaks are visually observed. We find the parameters estimates of the trend component using the robust Huber method with correlation coefficient $R = 0.42$. The model has the form as follows:

$$Traffic(t) = 1.498 \cdot 10^7 + 28645 \cdot t + X_1(t), \tag{10}$$

where $Traffic(t)$ is the observed values of the traffic at time t; $X_1(t)$ are model residues after subtracting the trend component from the original TS.

Standard deviation (SD) of the model is equal to $\varepsilon = 89.071$ MB, external SD is $\varepsilon_\Delta = 38.602$ MB.

To obtain an external standard deviation ε_Δ (forecasting errors), the initial TS is divided into two parts: 90% of the points are used to construct the model, and 10% – for the control (receipt ε_Δ).

The results of spectral analysis indicate the presence of periodic terms. The method of stepwise regression identified two statistically significant harmonics.

Harmonic model looks like:

$$X_1(t) = 6.694 \cdot 10^7 \cdot \sin\left(\frac{2\pi t}{24} + 245.51\right) + \\ + 3.323 \cdot 10^7 \sin\left(\frac{2\pi t}{4738} + 18.86\right) + X_2(t), \tag{11}$$

where $X_2(t)$ are residues after the subtraction of the harmonic model from $X_1(t)$.

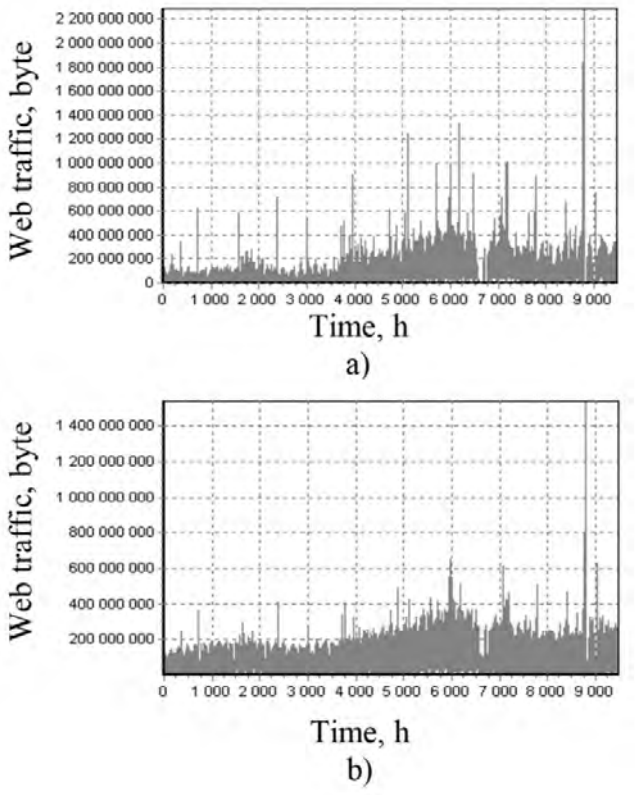


Fig. 1. a) Original TS diagram; b) Approximation of TS combined model diagram

Approximation accuracy of the model is 76.385 MB, external SD ε_Δ is 34.923 MB.

Identification of autoregressive-moving averaged (RAMP-approach) model are implemented by an algorithm that provides elimination of insignificant terms. A model of RAMP-approach $\{1, 0\}$:

$$X_2(t) = 0.629 \cdot X_2(t-1) + X_3(t), \tag{12}$$

where $X_3(t)$ are the residues of the model after subtraction of the AR component from $X_2(t)$.

SD of the model $\varepsilon = 59.356$ MB, $\varepsilon_\Delta = 31.616$ MB.

Engle test indicates the presence of conditional heteroscedasticity in the residues. Autoregressive conditional heteroscedasticity model RAMP-approach $\{1\}$:

$$X_3(t) = 0.263 \cdot X_3(t-1) + \varepsilon(t) + e(t), \tag{13}$$

where $\varepsilon(t)$ – the residues of the AR-model representing RAMP-approach $\{1\}$:

$$\varepsilon^2(t) = 2.0538 \cdot 10^{15} + 0.62203 \cdot \varepsilon^2(t-1), \tag{14}$$

$e(t)$ – the residues after the RAMP-approach.

As a result, the series is described by combined model including the trend, periodic components, RAMP-approach $\{1, 0\}$ and RAMP-approach $\{1\}$:

$$Traffic(t) = 1.498 \cdot 10^7 + 28645 \cdot t + \\ + 6.964 \cdot 10^7 \sin\left(\frac{2\pi t}{24} + 254.51\right), \tag{15}$$

where $\varepsilon(t)$ is given by (5), $e(t)$ are the residues of the model.

SD of the final combined model $\varepsilon = 55.385$ MB, external SD $\varepsilon_\Delta = 30.063$ MB.

The results of the last step residues diagnostics: residues are normally distributed; there is no autoregression.

IV. Discussions and Results: Interpretation of the Model Components

During the modeling of web traffic the increasing trend indicating that the information content increases from the beginning (September) to the end (May) of a school year is revealed.

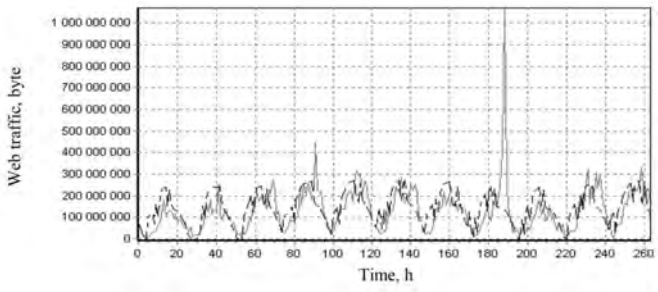


Fig. 2. Diagram of observation and predicted values of traffic to 264 hours

Two harmonics with periods of 24 hours and $6-6.5$ months are identified. Period of 24 hours indicates that the traffic increases gradually from the beginning of the day and during the first half of the day, reaching its peak by the middle of the day, then gradually decreases and by the end of the day takes its minimum value. $6 - 6.5$ months period characterizes the fact that in the summer holidays, winter holidays, and vacation the traffic is reduced, the active use of the site resources begins with the start of the semester, rising to its completion.

The presence of RAMP-approach effects is explained by variability (volatility) of the user accesses the site resources.

According to the model (15), we made the forecast at 264 hours. A satisfactory prediction interval (time interval at which

the predicted values correspond to the real data well) was found to be 48 hours. Fig. 2 shows a diagram of the predicted (dotted line) and original (solid line) values of TS at 264 hours.

Statistically significant correlation coefficient between the point estimates of forecast and original data for 264 hours is equal to 0.501, the range of satisfactory prediction is 0.758.

Comparison with the work of other researchers of the DPCGS has shown that the proposed models in the application of ARM-approach allows for a more accurate prediction of the parameters change.

The results obtained in the paper confirm the promise of using the so-called adaptive dynamic regressions being developed at the present time, for describing changes of DPCGS. Their advantages, compared with the traditional approaches to the analysis of time series, in particular, to the analysis of the variability of DPCGS, are: 1) expansion of the concept of the structure of the mathematical model describing the dynamics, 2) isolation of time-stable harmonics of oscillations, 3) several times increased accuracy of forecasting the changes on a certain time interval forward, which may have practical consequences.

V. Conclusion

It should be noted the initial assumptions of the regression analysis are always observed. However, discovering that the preconditions are violated is not sufficient. A specific software package containing particular measures that come into force under these conditions are required. Thus, for the effective use ARMP one should the apply a particular software package to automate the process of taking observations, analyze the quality of the models produced and analyze the compliance with the assumptions of regression analysis using the ordinary least squares method (LSM), as well as implement the appropriate procedures to adapt. The purpose of this study is to improve the performance of the computational modeling process by automating the search for the optimal set of regressors, and analyze it. To achieve this goal, it is necessary to solve a number of problems: 1) Development of the software package "Interactive Automated System for Optimal Regressions Modeling" (IASORM) based on connecting library quality analysis model with the compliance status of assumptions; 2) Implementation of the algorithm scenario of automatic data processing with the functional connection of libraries. The software package IASORM is a specialized system that implements the strategy of regression modeling. Automated script processing can improve the effectiveness of the existing methods of the package. Embedded library of the quality analysis and of the compliance model assumptions extend functionality for the user and is aimed at identifying the adequacy of models and observations in order to detect violations of the basic assumptions of regression analysis. Proposed scenario increases the computational process speed compared to interactive computing. IASORS implements the strategy of statistical (regression) modeling. This software package can be used to create regression models and predict dynamic processes.

A model of the dynamics of website server node is constructed. The combined model includes the components in the form of a trend, of two harmonics with periods of 24 hours and six months, ARMP approach models $\{1, 0\}$ and ARMP approach $\{1\}$. Decomposition of TS to systematic and random components allows for a more accurate interpretation of proper components.

The satisfactory prediction interval, which is 48 hours, is revealed.

Thus, knowing the forecast for a certain period, the site administrator will be able to estimate the available resources potential to address the upcoming workload and take appropriate action when it is increasing; for example, to increase the computational power of the server, to increase the number of servers, to create a cluster for parallel requests services, to install additional software that allows you to balance the workload, to speed up certain types of queries, etc.

The obtained results can be used for studies by regression estimation of space research data [6]–[8], astrophysics [9] and positional ground-based observations [10]–[12], as well as in geophysics [13].

Theoretically, the Huber method of estimators (in case of $1.5 < b$) and standard least square estimators are supposed to coincide in case of selection that obeys the normal distribution law. It is justified to reject the hypothesis of measurements equal accuracy. At the same time, it is fair to suggest the presence of some deviations from normal model of the measurement errors, in particular, the presence of abnormal errors. That means there are some abnormal measurements in the selection used for processing. The processing of traffic measurements, taken during a large time interval, with robust statistical procedure is one of the examples of its application [14]. The antijamming method has broad prospects: digital images processing; satellite geodetic measurements processing; data obtained by robotic systems processing; analysis of on-board telemetric data of aircraft (flight control) [15].

Acknowledgment

This work was partially supported by Russian Science Foundation, grants no. 20-12-00105 (according to the grant, the method for data analysis was created) and 19-72-00033 (according to the grant, the numerical calculations were carried out). This work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University. This work was partially supported by a scholarship of the President of the Russian Federation to young scientists and post-graduate students SP-3225.2018.3, the Russian Foundation for Basic Research grant no. 19-32-90024 and the Foundation for the Advancement of Theoretical Physics and Mathematics "BASIS".

References

[1] S. G. Valeev, "Regression modelling in selenodesy," Earth Moon Planets, 1986, vol. 35, iss. 1, pp. 1–5.

[2] P. Huber, B. Kleiner, T. Gasser and G. Dumermuth, "Statistical methods for investigating phase relations in stationary stochastic processes," IEEE Trans. Audio Electroacoust., 1971, vol. 19, iss. 1, pp. 78–86.

[3] X. Zong, Q. Sun, D. Yao, W. Du and Y. Tang, "Trajectory planning in 3D dynamic environment with non-cooperative agents via fast marching and Bézier curve," Cyber-Physical Systems, 2019, vol. 5, iss. 2, pp. 119–143.
[4] Y. A. Nefedyev, S. G. Valeev, R. R. Mikeev, A. O. Andreev and N. Y. Varaksina, "Analysis of data of "Clementine" and "KAGUYA" missions and "ULCN" and "KSC-1162" catalogues," Adv. Space Res., 2012, vol. 50, iss. 11, pp. 1564–1569.
[5] L. V. Rozovskii, "Accuracy of the normal approximation," J. Sov. Math., 1992, vol. 61, iss. 1, pp. 1911–1918.
[6] I. Davoodabadi, A. A. Ramezani, M. Mahmoodi-k and P. Ahmadizadeh, "Identification of tire forces using Dual Unscented Kalman Filter algorithm," Nonlinear Dyn., 2014, vol. 78, iss. 3, pp. 1907–1919.
[7] W. Chen and R. Tenzer, "Harmonic coefficients of the Earth's spectral crustal model 180ESCM180," Earth Sci. Inform., 2015, vol. 8, iss. 1, pp. 147–159.
[8] Y. A. Nefedjev and A. I. Nefedjeva, "Determination of refraction anomalies by global inclinations of airstratas of identical density," Astron. Nachr., 2005, vol. 326, iss. 8, pp. 773–776.
[9] L. Monostori, "Cyber-physical production systems: Roots, expectations and R&D challenges," Procedia Cirp, 2014, vol. 17, pp. 9–13.
[10] M. G. Sokolova, Y. A. Nefedyev and Varaksina N. Y. "Asteroid and comet hazard: Identification problem of observed space objects with the parental bodies," Adv. Space Res., 2014, vol. 54, iss. 11, pp. 2415–2418.
[11] T. B. Lieu Tran, M. Törngren, H. D. Nguyen, R. Paulen, N. W. Gleason and T. H. Duong, "Trends in preparing cyber-physical systems engineers," Cyber-Physical Systems, 2019, vol. 5, iss. 2, pp. 65–91.
[12] Y. A. Nefedjev and N. G. Rizvanov, "The results of an accurate analysis of EAO charts of the Moon marginal zone constructed on the basis of lunar occultations," Astron. Nachr., 2002, vol. 323, iss. 2, pp. 135–138.
[13] V. V. Lapaeva, V. P. Meregin and Y. A. Nefedjev, "Study of the local fluctuations of the Earth's crust using data of latitude observations," Geophys. Res. Lett., 2005, vol. 32, iss. 24, p. L24304.
[14] S. G. Valeev, "Coordinates of the Moon reverse side sector objects," Earth Moon Planets, 1986, vol. 34, iss. 3, pp. 251–271.
[15] M. Elmi, H. A. Talebi and M. B. Menhaj, "Robust adaptive dynamic surface control of nonlinear time-varying systems in strict-feedback form," Int. J. Control. Autom. Syst., 2019, vol. 17, iss. 6, pp. 1432–1444.

Cryptographic Algorithm Implementation for Data Encryption in DBMS MS SQL Server

Olga A. Safaryan
Information System Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
safari_2006@mail.ru

Evgenia V. Roshchina
Information System Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
ev_roschina@mail.ru

Larissa V. Cherckesova
Information System Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
chia2002@inbox.ru

Elena V. Pinevich
Information System Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
hpinevich@mail.ru

Andrey G. Lobodenko
Information Systems and Radioengineering Department
Shakhty Branch of
Don State Technical University
Rostov–on–Don, Russia
andrey@sssu.ru

Boris A. Akishin
Mathematics and Computer Science Department
Don State Technical University
Rostov–on–Don, Russia
akiboralex@mail.ru

***Abstract*— the report discusses issues related to the MS SQL database administration, as well as the basic methods for implementing backup and encryption. It is proposed to create an application that backs up databases in the MS SQL Server DBMS using the GOST R 34.12 "Kuznechik" (Grasshopper) encryption algorithm. Developing the application, which provides the secure backups it is necessary to study MS SQL VDI technology and to implement this interface in the application to transfer the backup stream directly to the application under development.**

Keywords— database management system, structured query language, SQL–Server, Virtual Device Interface, database backup, encryption methods, cryptographic algorithm

I. INTRODUCTION

The backing up databases is important aspect of security in enterprises using servers with database management systems. There are many ways to implement secure backups. One of this way is to create encrypted backup using encryption algorithms. DBMS MS SQL Server provides several encryption algorithms available to the user, among them are encryption standards such as "AES" (Advanced Encryption Standard), "DES" (Data Encryption Standard) and "TDES" (Triple Data Encryption Standard, 3DES). Russian encryption standard is GOST R 34.12, the latest edition of which was in 2018. But there is revision slightly older than 2015, in its case the cryptographic strength (security) and complexity of encryption algorithms have already been analyzed, and according to experts' analysis, this revision is considered to be very reliable [1], [2].

Report research object is backing up databases; research subject is database backup methods using encryption in MS SQL Server Data Base Managing System. Research purpose is studying the possibility of using GOST R 34.12 "Kuznechik" (Grasshopper) algorithm instead of the AES algorithm when backing up databases in MS SQL Server DBMS; and to evaluate the resulting import–substitution crypto algorithm effectiveness.

Developing the application, which provides secure backups it is necessary to get acquainted with MS SQL VDI (Virtual Device Interface) technology and implement this interface in the application to transfer the backup stream directly to the application under development. This algorithm involves using of Russian standard in conjunction with MS SQL VDI technology, which allows transferring the stream of the created backup directly to developed software, which eliminates the possibility of leaking a copy of database before it is encrypted.

II. BASIC PROVISIONS

Database (DB) is a structured repository with information. Databases cannot exist fully by themselves; it is just a set of information in its original form. A database management system (DBMS) is required in order to manipulate information. By database management, it means the possibility of individual or collective editing of information, its sorting, full or partial copying and moving, as well as the ability to combine several databases into one common system.

The DBMS software tool products were created for implementing the previously listed functions. The main components of the DBMS are presented in Fig 1.

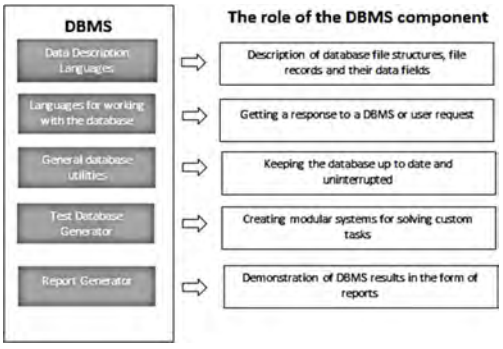


Fig. 1. DBMS components

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

Database Management System (DBMS) is specialized software having main task is to ensure uninterrupted access to the database, as well as to the methods of its administration.

An extremely important aspect of the database smooth operation is the DBMS maintenance, which is done by database administrators. Database Administrator is a person responsible for the database requirements development, its design, implementation, effective use and maintenance.

Most DBMSs have the ability to back up the database. Creating the backup copy is necessary to restore the database to stable state in case of failures and errors. Backup is carried out by the DBMS itself, it can be performed manually by the administrator, or it can be scheduled and performed automatically in the specified time interval.

A backup consists of the database and all the log files required to bring the database copy to consistent state during the backup. It should be noted that regardless of the backup type being performed, the backup data must be intact.

Therefore, before performing the backup, the user should check the database stability and perform this backup procedure only when the database is in some stable state.

Microsoft SQL Server is relational database management system (RDBMS) developed by Microsoft Corporation. The main language used is Transact–SQL. This is relational distributed client–server DBMS. The main tool is SQL Server Management Studio (SSMS). It is a graphical user interface (GUI) and Transact–SQL scripting interface for managing the database engine component and databases.

Microsoft SQL Server has all the necessary tools to maintain a stable database. It is important to understand that SQL Server is not single monolithic application but is structured as a series of components. The components of SQL Server and their purpose are shown in the Table 1. The main components are Database Engine, SSAS, SSIS and SSRS.

Transact–SQL (T–SQL) is Modified SQL language. All applications interacting with an instance of MS SQL Server, regardless of their implementation and user interface, send Transact–SQL statements to the server.

One of the most important aspects is to ensure that data is backed up regularly so that if a failure occurs, the database can be restored. Despite the fact that the computer industry has been aware of the need for reliable backup strategies for decades, tragic histories of data loss are still commonplace.

SQL Server recovery models: SQL server supports 3 types of database recovery models. All models store data in event of accident, but there are important differences that administrator need to know when choosing recovery model for database [1]. The recovery models are described in Table 2. Each database can have its own recovery model. The model is selected in the Database Properties window on the Parameters page.

Backing up databases and logs backup types SQL Server supports several types of backups, that administrator can combine to implement the right backup and recovery strategies for specific database based on business requirements and recovery goals. Description of backup types is given in Table 3.

The main advantage of encrypting backups in MS SQL Server is that the process itself occurs directly when creating a backup. Thus, the backup file is saved to disk already in encrypted form. In this case, a high speed of creating the encrypted backup is ensured, since encryption occurs on the fly, and the system integrity is not violated.

The main disadvantage of MS SQL Server encryption is that it is not possible to expand the available arsenal of ciphers. The MS SQL Server software product policy does not imply an extension of the available cipher suite. The user must choose among the available options [3–5].

TABLE I. MS SQL COMPONENTS

Component	Description
Database Engine	The core of the SQL Server platform. Provides high performance and scalability of relational databases based on the SQL language. It can be used to place data, for further processing using transactions online and to create data warehouses. SQL Server also includes optimized database engine that leverages inmemory technology to improve the performance of short transactions.
Analysis Services	SQL Server Analysis Services (SSAS) – Analysis Services, provides OLAP (Online Analytical Processing) and data analysis functionality for business intelligence applications. They allow the organization to collect data from several sources, for example, relational databases, and process them in various ways.
Integration Services	SQL Server Integration Services (SSIS) – Integration services, the enterprise–wide tool for extracting, transforming, and integrating data from various sources and moving them to one or more target data sources. Designed to merge data from heterogeneous sources and load them into storage, data marts, etc.
Reporting Services	SQL Server Reporting Services (SSRS) – Reporting Services, includes Report Manager and Report Server. They are the full–blown server platform for creating, managing and distributing reports. Allows administrator to use combination of SQL Server and IIS capabilities to process and store reports. SSRS can be installed independently or integrated with Microsoft SharePoint.
Master Data Services	SQL Server Master Data Services (MDS) – the environment for creating business rules that guarantee the quality and accuracy of master data. Business rules can be used to run business processes that perform checks and control data flows in the database.
Data Quality Services	SQL Server Data Quality Services (DQS) – environment for creating a knowledge base metadata repository that improves the quality of organization data. Data cleaning processes allow administrator to modify or delete incomplete and incorrect data, matching processes allow administrator to identify and combine duplicate data.
StreamInsight	SQL Server StreamInsight provides the platform for creating applications that handle complex events for real–time data streams.
Full–Text Search	Full–text search is Database Engine feature that provides the sophisticated semantic search engine for text data.
Replication	The Database Engine includes replication, a set of technologies for synchronizing data between servers to meet the needs of data distribution
Power View for SharePoint Server	Power View is component of SQL Server Reporting Services. Provides interactive data research, visualization and presentation. Power View is also available in Excel.

As in any field, there are alternatives. If encrypted backups are created, all available options are divided into three groups. Each of these groups has its own advantages and disadvantages, for a complete understanding it is necessary to consider each group in more detail.

TABLE II. DATABASE RECOVERY MODELS

Recovery model	Description
Simple model	Designed to restore to the last backup point. The strategy for this model should include full and differential backup operations. By enabling a simple recovery model, administrator cannot back up transaction logs. Automatically trimming the log at the checkpoint (clearing all inactive transaction records) to minimize space. This model is ideal for most system databases.
Full model	Designed to restore the database to the point of failure or at a specific point in time. When using this model, all operations are logged, including bulk operations and bulk data loading. The strategy should include the following archives: full, differential and transaction log archives (or only full archives and transaction log archives), i.e. log backup required. Eliminates data loss due to a damaged or missing data file.
Bulk Logged model	This model reduces the space occupied by the transaction log while retaining most of the functionality of the full recovery model. Minimal logging of bulk operations and bulk downloads is performed without controlling individual operations, which can improve the performance of bulk copy operations. If the failure occurs before full or differential backups are performed, bulk operations and bulk downloads will need to be repeated manually. The backup strategy for this model should include the same archives as for the full model.

The first group should include third–party software tool that aims to simplify the administration of MS SQL Server. In this case, the application independently sends queries to the MS SQL server and configures the task *scheduler*.

In turn, the user operates with an intuitive program interface, where parameters and settings are provided usually in simplified version. Programs such as "*Effector saver"* and "*HandyBackup"* provide the user with all the necessary tools for creating backups, scheduling, and restoring the database.

For their part, the programs it selves do not have their own set of encryption algorithms, but only use what MS SQL Server offers, in which case, administrator must first configure encryption on the *MS SQL server* itself. In addition, there is no way to expand the list of available algorithms with other algorithms. The second group includes whole cluster of applications that allow administrator to encrypt an arbitrary data set, including the finished backup file of *MS SQL Server.*

Programs such as "*Files Cipher"* and *"dsCrypt"* have many cryptographic algorithms and allow administrator to encrypt an arbitrary set of files.

As the main drawback, it should be noted that for encryption, the file must be created already physically, and located in the directory accessible to programs. Therefore, such encryption is vulnerable. In the case where attacker has access to the memory section where the files are stored in the clear, he can grab the file before encrypting. In the case of using such programs for encrypting backups, main vulnerability will be that backup file in unencrypted form will be created previously [6].

The third group includes applications that allow administrator to encrypt a hard disk partition, thus encrypting, on the fly, all data that is written to this disk. Programs such as "*Truecrypt*" and "*BestCrypt*" have enough large set of crypto algorithms, but, like in the case of all the other programs listed above, the Russian import–substitution GOST is not included in the list of cryptographic algorithms.

TABLE III. DESCRIPTION OF BACKUP TYPES IN SQL SERVER

Backup type	Description
Full	Full database backup includes all objects, system tables and data, as well as fragments of transaction logs corresponding to the backup time. A full backup allows administrator to fully restore the database at the time the backup is completed.
Differential	Differential backups are designed to archive data that has changed since the last backup. When backing up, only changes are saved, so archiving takes less time and can be done more often. Differential copies also include fragments of transaction logs needed to restore the database after the backup is completed.
Transaction Log	When a transaction log is backed up, the archive saves the changes that have occurred since the last backup of the transaction log, and then truncates it, during which completed or canceled transactions are erased. Unlike full or differential archives, the transaction log archive records the log state at the time the backup operation started (and not at the time, when it was completed).
File / File Group	These backups do not allow archiving the entire database, but only the specified files and file groups. Backups of this type are used when working with large databases to save time. Archiving files and filegroups has the number of features. At the same time, the transaction log must also be archived. Therefore, this method does NOT perform archiving if the Truncate Log On Checkpoint option is enabled. If database objects span multiple files or filegroups, administrator must back up ALL of these files or filegroups.
Partial	It contains data from only some filegroups of the database, including data from the primary filegroup, all filegroups that are read / write, as well as any additional specified files that are read–only. Partial backups can be useful when administrator need to exclude read–only filegroups. A partial backup of read–only database contains only the primary filegroup.
Tail–log Backup	The backup copy of the final log fragment contains all the records whose backup has not yet been created (the final log fragment, the "tail"), which helps to prevent loss of work and keep the chain of logs intact. To restore the SQL Server database to the last moment in time, administrator must first back up the final fragment of its transaction log. The final piece of the log becomes the last part of the backup in question in terms of restoring the database.
Copy Only	Dedicated backup that is independent of the regular SQL Server backup sequence. Typically, creating the backup changes the database and affects how subsequent backups are restored. However, sometimes administrator have to back up a database for special needs when it does not affect the overall backup and recovery process. For this purpose, backup copies are used only for copying (databases or transaction logs).

It should be noted, that backup entry to encrypted volume of the hard disk does not imply that the file will be encrypted. It is also worth paying attention to the fact that in a case of encryption of hard disk volume, the performance will be reduced, which may cause additional inconvenience in working.

Basic methods for implementing database backup, as well as its encryption using built–in MS SQL tools, are considered. Besides, analysis of alternative solutions and their capabilities for creating backups with subsequent encryption was carried out.

Nowadays when processor power and memory transfer speed are increasing every year, standards released several years ago can become out of date very quickly.

In the case of cryptographic algorithms, standards are designed usually with a margin of power and reliability, thus providing stable protection for several decades. However, despite the expected margin, often during the operation of the developed standards, vulnerabilities are discovered that can greatly reduce the expected life of the standard. Example is the "DES" standard, the estimated service life of which, according to experts, was supposed to last until the middle of XXI century, but, in practice, the life was much shorter. "DES" was national standard in United States and existed as such to 1980. However, in "DES" standard many vulnerabilities to multiple attacks were found, and it was replaced by "TDES".

The standards popularity should be noted also. As better the standard is known, the more software solutions are based on it, the more attacks are made on it, thus significantly speeding up the search for possible vulnerabilities and shortening the standard's estimated reliability period [7].

III. Algorithmic construction

The program application under development should be atomic for the user. Customer does not need to know about the business logic that processes his requests.

After the user's request, the application should respond with the result, regardless of success. In addition, the application must be fault tolerant, as there is always the possibility that the user will indicate incorrect data, perform actions in the wrong order, or the program will crash.

The application under development will back up the database, as well as restore it from backups. In addition, backup copies must be encrypted, this is necessary for their safe storage on the user's physical disk.

Algorithm "Kuznechik" (Grasshopper) GOST R 34.12–2015 was selected as encryption algorithm. It is also necessary to keep records of user actions, and the application must have its own database for keeping records and storing user data.

The portable database system MSQLite 3.0 can serve as such database. The main problem that administrator will have to face at the development stage is that MS SQL only supports its own encryption out of the box and does not provide to the third–party developers with the opportunity to expand the available set of algorithms. At the same time, the Microsoft development company has provided an alternative option, in which case third–party developers are provided with VDI interface through which administrator can redirect the backup stream directly to the third–party application. Thus, the created database will be redirected to application being developed. The final scheme of developed application is demonstrated in the Fig 2.

Based on the presented diagram it can be seen that the user interacts exclusively with the application being developed. Also, the user should be granted access to the event log, in which, in chronological order, all interactions with databases on *MS SQL Server* will be recorded [8], [9].

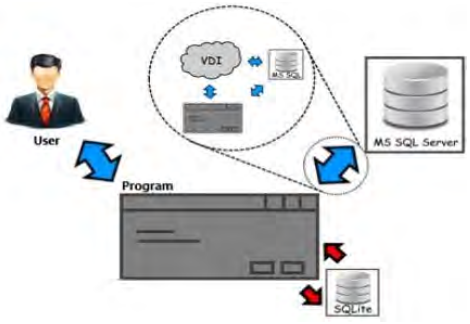


Fig. 2. Architecture diagram of developed application

One of important functions of the developed software tool is backup copy creating. The function that implements the backup mechanism must connect to the *MS SQL Server* in order to initialize the *VDI interface* and for the subsequent transfer of the database backup stream. The parameters for receiving must be stored in the program settings to automate the process and simplify the work with the server. Resulting stream can be encrypted with "Kuznechik" (Grasshopper) import–substitution algorithm. The user also indicates the path to save the backup in the program settings. In addition, all program actions with the server must be documented in the event log of the latest program activities. Data array encryption method:. Thus, it will be possible to restore the chronology.

```
public byte[] Encript(byte[] file, byte[] masterKey) masterKey
  = Encoding.Default.GetBytes (LengthTo32Bytes
  (Encoding.Default.GetString(masterKey)));
  KuzKeyGen(masterKey);
int NumOfBlocks;
int NumberOfNull;
byte[] OriginByteArray = file;
byte[] encypted = new byte[0];
if ((file.Length % 16) == 0)
NumOfBlocks = file.Length / 16;
Array.Resize(ref encypted, file.Length);
else NumOfBlocks = (file.Length /16) + 1;
NumberOfNull = NumOfBlocks*16 – file.Length;
int StartLength = file.Length;
Array.Resize(ref OriginByteArray, OriginByteArray.Length +
                                                NumberOfNull);
Array.Resize(ref encypted, OriginByteArray.Length);
if (NumberOfNull ==1 ) OriginByteArray[OriginByteArray.
Length – 1] = 0 × 80;
else –
for (int i = OriginByteArray.Length – 1; i >= 0; i--)
if (i == OriginByteArray.Length - 1)
OriginByteArray[OriginByteArray.Length – 1] = 0x81;
else if (OriginByteArray[i] ! = 0)
OriginByteArray[i + 1] = 0x01; break;
for (int i = 0; i < NumOfBlocks; i++)
byte[] byteBlock = new byte[16];
for (int j = 0; j < 16; j++)
byteBlock[j] = OriginByteArray[i * 16 + j];
for(int j = 0; j < 9; j++)
byteBlock = KuzX(byteBlock, iterK[j]);
byteBlock = KuzS(byteBlock);
byteBlock = KuzL(byteBlock);
byteBlock = KuzX(byteBlock, iterK[9]);
for (int j = 0; j < 16; j++)
```

```
encypted[i*16 + j] = byteBlock[j];
return encypted;
```

Creating the backup with next encryption:

```
public async void CreateBackUpWithEncryption(string
fullFileName, string kuzKeyPath)
await Task.Run(() =>
MainForm.AppendTextRow("Started backup with encryption.");
var key = KuzKey.ReadKeyFromFile(kuzKeyPath);
MainForm.AppendTextRow("Key loaded.");
if (!string.IsNullOrWhiteSpace(key.Error))
MainForm.AppendTextRow("No errors were detected in the key.");
else MainForm.AppendTextRow(key.Error );
var keyByteArray = key.GetCipherKey();
string commandToCreate = string.Format(@"backup -s {0} -d {1}
--stdout -u {2} -p {3}",
DataBaseData.LocalUserSettings.ServerName,
DataBaseData.LocalUserSettings.DataBaseName,
DataBaseData.LocalUserSettings.UserName,
DataBaseData.LocalUserSettings.Password);
var process = new Process StartInfo = new ProcessStartInfo
UseShellExecute = false, CreateNoWindow = true,
RedirectStandardOutput = true, RedirectStandardError = true,
RedirectStandardInput = true, FileName = @"PackDb.exe",
Arguments = commandToCreate EnableRaisingEvents = true try
Ghost2015 cipher = new Ghost2015(); process.Start();
MainForm.AppendTextRow("Launch VDI." );
using (MemoryStream stream = new
MemoryStream())MainForm.AppendTextRow("Creating a
backup."); process.StandardOutput.BaseStream.CopyTo(stream);
MainForm.AppendTextRow("Transferring a stream to cryptographic
algorithm.");
File.WriteAllBytes(fullFileName, cipher.Encript(stream.ToArray(),
keyByteArray)); MainForm.AppendTextRow("Encryption
completed." ); process.WaitForExit();
MainForm.AppendTextRow("The specified path backup was created.");
MainForm.AppendTextRow(" ---------- --." );
writeEncryptedBackUpInfoToDataBase(fullFileName, kuzKeyPath);
catch (Exception ex)
{throw ex;}
```

In case of errors in user settings, failures while connecting to the server, or errors while receiving the backup stream, the algorithm is interrupted and notifies the user of malfunction.

This part of the algorithm should take into account all possible exceptions and process them in the correct form in order to convey as accurately as possible to the user the nature of the error that has occurred [10].

The work result must be recorded in the event log before the algorithm completion. This is necessary so that the user can view the results of the work, as well as, if necessary, find out what parameters were backed up.

The algorithm shown in the Fig. 3 is a backward backup algorithm. It should be noted, that regardless of whether the backup was encrypted or not, the final steps to restore the database occur directly on the MS SQL server.

The work success, as in the case with the backup algorithm, should also be recorded in the event log upon exiting the algorithm. The parameters are received from the user at the entrance to the algorithm. These parameters include settings for connecting to MS SQL Server. In the case of encrypting the backup, the key parameters are received. During execution, the algorithm may be interrupted by an error; in this case, the algorithm is exited, and the user is notified of the error.

In the case of backups, keeping track of actions is important part of the entire process. Since it is necessary to know the result of performing actions with the database, so that in case of failures, you can determine the nature of the error. To do this, the DBMS has a log that records all user actions, as well as records of failures with their detailed description.

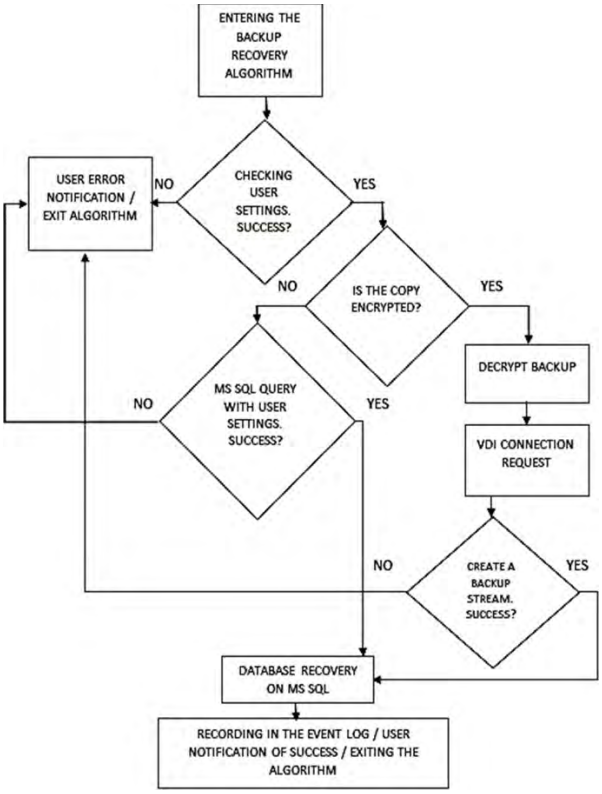


Fig. 3. Restoring Database from Backup

In the case of backups, keeping records of activities is an important part of the whole process. It is necessary to know the result of performing actions with the database in the failure case because it would be possible to determine the nature of the error.

To do this, there is log in the database management system (DBMS), in which all user actions are recorded, as well as failures are recorded with their detailed description.

In the software product being developed, the event log is collection of tables in an embedded database. The structure of these tables is shown in the Fig. 4.

It should be noted that the database table entry is carried out only if successful. Thus, looking at event log, administrator can check the operation result. The tables are interconnected by encryption key ID. In this case, one–to–many relationships can be traced, since many backups can be encrypted with same prickle, but backup copy is encrypted with only one specific key. The key table stores information about key maker and creation

date. Name of creator in this case corresponds to login from MS SQL Server, on behalf of which user creates the key.

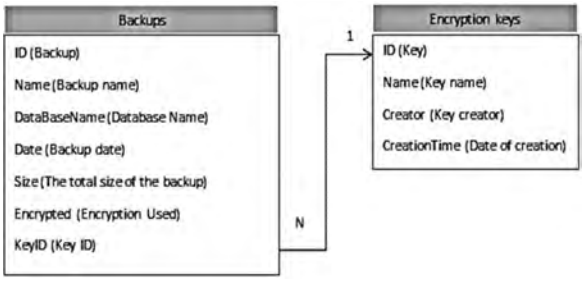


Fig. 4. Event Log Tables

The main problem that can be encountered during the encrypted backup creation is the transfer of the backup file itself to the program being developed. Since MS SQL Server does not allow expanding the set of ciphers offered to the user, the developed program must encrypt the backup copy after its creation. In this case, before transferring the backup file, it will be created on the media, and at this point in time, it will be available to the intruder [11], [12]. To rid developed software product of such vulnerability, it is necessary to use VDI interfaces, through which DBMS will transfer the stream of created backup directly to developed program tool.

VDI (Virtual Device Interface) – SQL Server provides an API called Virtual Backup Device Interface (VDI), which allows third-party software vendors to integrate SQL Server into their products to support backup and recovery operations. Designed for maximum reliability and performance, these APIs also support the full range of SQL Server backup and recovery features, including a full range of quick snapshot and backup capabilities.

Thus, creation of backup file on physical medium in unencrypted form will be excluded. If transfer of backup stream is completed, connection via VDI interface ends automatically.

IV. Conclusion

This report presented algorithms for creating an encrypted backup, as well as restoring a database from an encrypted backup. These algorithms involve the use of the Russian standard GOST R 34.12 “Kuznechik” (Grasshopper) in the conjunction with MS SQL VDI technology, which allows administrator to transfer the stream of the backup created directly to the developed software tool, which eliminates the possibility of database copy leak before the encryption. The realized import–substitution software allows administrator to protect reliably database backups in MS SQL Server.

Thus, the investigation aim is achieved. After studying the possibility of application of GOST R 34.12 “Kuznechik” (Grasshopper) algorithm instead of the “AES” algorithm when backing up databases in the MS SQL Server DBMS, the following results were obtained. The comparative analysis showed that the encryption algorithms "AES" and "Kuznechik" are similar in many ways; both are based on the "SP–network" (substitution–permutation network is the type of block cipher proposed in 1971 by H. Feistel) and support 265–bit key. However, at the same time, "AES" algorithm appeared much earlier, is also more well–known and popular encryption algorithm in the world, and is widely used in software products, including Microsoft products.

This fact is its vulnerability, since the number of attacks on the "AES" algorithm is tremendous, and significantly exceeds the number of attacks on "Kuznechik"(Grasshopper) algorithm, which gives more possibilities, that an attacker will find the vulnerability in this cipher. It should be noted also, that the "Grasshopper" algorithm is much easier to understand and implement, requires fewer computing resources and has more operating speed and performance.

Evaluating the effectiveness of the resulting crypto algorithm shows that replacing “AES “with "Kuznechik" improves the performance of developed software application. The implementation of crypto algorithm was tested and verified for compliance with its formal model theoretically, and in the laboratory conditions available to authors.

At the same time, MS SQL Server DBMS is commercial product, and getting access to its programming code for the research of the functional, when replacing the AES encryption algorithm with import–substitution GOST R 34.12 "Kuznechik" (Grasshopper) is quite problematic.

However, theoretical studies show that this replacement leads to significantly faster operation speed and performance of the developed software application and allows increasing the efficiency of the import–substitution algorithm and corresponding software by about 20%.

References

[1] A.K. Otinchiev, “Using Dapper C # in programming”. – SPb.: Peter. 2019. – 386 p. (In Russian).

[2] GOST R 34.12–2015. “Cryptographic Information Protection. Block Ciphers. Code "Kuznechik” (Grasshopper) and the modes of operation of block ciphers”. Introduced. 2015–07–19. –M.: Publishing house of standards, 2015. – 21 p. (In Russian).

[3] A.V. Cheremushkin, “Cryptographic protocols: basic properties and vulnerabilities”. Institute of Cryptography. Communications and Informatics. – M. Stringer, 2009. – 116 p. (In Russian).

[4] GOST 28147–89, “Information processing systems. Cryptographic protection.Cryptographic conversion algorithm”. Introduced. 1990– 03–05. –M.: Publishing house of standards, 1989. – 35 p. (In Russian).

[5] W.R. Stanek, “Microsoft SQL Server 2008. Administrator's Pocket Consultant”. Microsoft Press. 2010. – 738 p.

[6] E.K. Baranova, Yu.A. Rodichev, “Information security and protection”. – M.: RIOR, Infra–M, 2017. – 324 p. (In Russian).

[7] Federal Information Processing Standards Publication 1997. Announcing the Advanced Encrytpion Standard (AES), Nov 2001.

[8] E. Shmuelia, R. Vaisenberg, E. Gudesc, etc. “Implementing database encryption solution, design and implementation issues” Computers & Security. Vol. 44. July 2014, Pp. 33–50.

[9] “Securing Network Services and Protocols MCSE. Study Guide Designing security for Windows Server”. Chapter 5. 2003. Network: Exam 70–298. 2004. Pp. 241–350.

[10] D. Cherry, “Database Backup Security Securing SQL Server (3rd Edition) Protecting administrator Database from Attackers”. 2015. Chapter 10. Pp. 293–311.

[11] S. Nakamura, C. Qian, S. Fukumoto, etc. “Optimal backup policy for database system with incremental and full backups”. Mathematical and Computer Modelling. Volume 38. Issues 11– 13, December 2003, Pp. 1373–1379.

[12] M. Hatzopoulos and J. Kollias, “Optimal policy for database backup and recovery”. Information Processing Letters Volume 12, Issue 2, 13 April. 1981. Pages 55–58.

Increasing Self-Timed Circuit Soft Error Tolerance

Igor Sokolov
Institute of Informatics Problems
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Moscow, Russia
ISokolov@ipiran.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 Diachenko
Institute of Informatics Problems
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Moscow, Russia
diaura@mail.ru

Yury Rogdestvenski
Institute of Informatics Problems
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Moscow, Russia
YRogdest@ipiran.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

***Abstract*—Indication subcircuit is an essential part of the self-timed circuits. It provides acknowledgment of the self-timed circuit switching completion and ensures correct handshake interaction between functional blocks. Besides, indication subcircuit complexity is comparable with the indicated self-timed circuit's complexity. So short-term soft errors, induced by the external and internal causes in both the indication subcircuit and the indicated self-timed circuit, are equally dangerous. Indication subcircuit soft error tolerance depends, the first, on its immunity to soft errors in the indicated self-timed circuit and, the second, on its failure protection. The first aspect becomes lower critical due to the XOR cell on the first stage of the indication subcircuit. An appropriate circuitry basis decreases indication subcircuit sensitivity to the possible soft errors induced in it. Static and semi-static Muller's C-element is a traditional base component used for indication purposes. Its dual interlocked implementation improves the indication subcircuit failure protection against soft errors in its internal nodes, but not sufficiently. The article proposes a new C-element's schematic that fully tolerates it against the soft errors in all internal nodes. Besides, using C-elements with in-phase inputs and output in an indication pyramid ensures indication subcircuit protection against soft errors induced at the output of the C-elements. The proposed approach makes an indication subcircuit fully protected against all soft errors induced in it.**

***Keywords*—C-element, DICE-style, dual-rail code, indication, self-timed circuit, soft error, spacer, tolerance, working state**

I. Introduction

Studies prove that combinational self-timed (ST) circuits are naturally immune to short-term soft error upsets and transients (SEUT) [1] induced by ionization effects because of single nuclear particles and cosmic rays [2, 3]. This feature is due to dual-rail signal discipline and appropriate layout placement of the cells. An indicator subcircuit, an integral part of the ST circuit, operates with unary phase signals that indicate dual-rail information signals. It is also susceptible to adverse factors causing soft errors.

An indication subcircuit of any ST circuit collects all partial indication signals into one total output acknowledging the entire ST circuit switching to each operation phase. The indication subcircuit layout occupies up to a half part of the ST circuit die area, and SEUT can appear in the indication subcircuit with a probability close to the probability of the same event in another part of the ST circuit.

Therefore, the estimation of the ST circuit's indication subcircuit tolerance to SEUTs and the development of methods improving this tolerance is an urgent task. The paper analyzes critical SEUT in the indication subcircuit of the ST circuits implemented in a 65-nm and below CMOS bulk process. It proposes circuit-based methods increasing its tolerance to SEUTs.

The scientific novelty of this article consists of two ideas.

The first idea is to use a four-transistor converter in the Dual Interlocked Cell (DICE) C-element instead of a conventional two-transistor one. This replacement prevents any C-element output changes because of the short-term SEUT at its inputs and internal nodes.

Another idea is to build the ST circuit's indication subcircuit using C-elements with two in-phase outputs to improve indication subcircuit tolerance to soft errors.

The results of simulating both the proposed and known DICE-like C-elements with inserted soft error sources and a soft error hardened indication subcircuit structure are the main contribution of this paper.

II. Indication Subcircuit Basis

The dual-rail signal consists of two components coding one information bit. It has two working states ("01" and "10") and one spacer ("00" for null spacer or "11" for unit spacer).

This work was supported in part by Grant of the Ministry of Education and Science of the Russian Federation for Application No. 2020-1902-01-016 in FRC CSC RAS.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

Remain binary state opposite to the spacer ("11" for null spacer or "00" for unit spacer) is prohibited.

Respectively, classical dual-rail signal indication [4] bases on detecting any dual-rail signal component transient from the spacer to a working value and vice versa and thus simplifies the indication implementation. For example, NOR2 cell indicates a dual-rail signal with a zero spacer, whereas cell NAND2 indicates a dual-rail signal with a unit spacer. Such an implementation assumes that the anti-spacer never appears. As a result, an indication subcircuit considers the anti-spacer caused by a SEUT as a working state. It can lead to disruption of the ST circuit operation by corrupting processed data or to dead-lock of a handshake between ST circuit parts. Such errors are a significant part of the critical failures in ST circuits.

The circuitry method proposed in [1], based on the anti-spacer indication as the second spacer using XOR or XNOR cells, masks the anti-spacer and significantly increases a combinational ST circuit tolerance to short-term SEUT. Fig. 1 shows two XOR implementations suitable for use in ST circuits. The circuit in Fig. 1(a) is convenient for implementation with a standard cell library. The circuit in Fig. 1(b) has the least possible complexity.

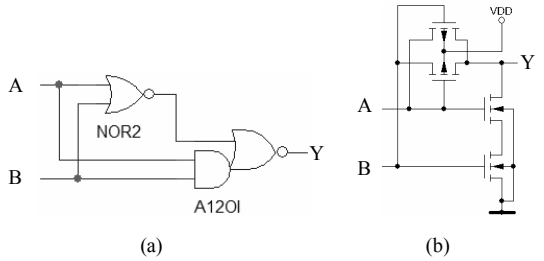


Fig. 1. XOR implementations: a) on standard cells; b) on CMOS transistors

Analysis shows [1] that the use of XOR as the first stage of the indication subcircuits masks soft errors caused by the anti-spacer in ST combinational circuits and increases their fault tolerance.

The problem of indication subcircuit SEUT tolerance has two aspects. The first is indication subcircuit immunity to SEUT in the indicated ST circuit, and the second is protection against SEUT that appears in the indication subcircuit. SEUT in the combinational ST circuit causes any of the following three SEUT situation types:

- dual-rail signal's incorrect working state,
- dual-rail signal's premature spacer,
- dual-rail signal's anti-spacer.

Indication subcircuit cannot recognize the first two events as failures because they are valid in the dual-rail signal encoding. The XOR and XNOR cells mask the third situation considering it to be a dual-rail signal spacer state [1]. Therefore, using XOR and XNOR cells as first-level dual-rail signal indicators makes indication subcircuit immune to 33% of the possible SEUT types in the indicated ST circuit.

The indication subcircuit circuitry basis determines the indication subcircuit tolerance to SEUT, induced directly in it. An indicator collecting partial indication signals into a single total signal operates by the following Muller's element function [4]:

$$O^+ = I_1 \cdot I_2 \cdot \ldots \cdot I_N + O \cdot (I_1 + I_2 + \ldots + I_N), \qquad (1)$$

where O and O^+ are the current and next values of the total indication signal respectively; I_1, I_2, …, I_N are partial indication signals. Fig. 2 illustrates a typical function (1) static implementation on CMOS transistors, while Fig. 3 shows its semi-static CMOS implementation. Here, transistors Mn and Mp gated by feedback signal ensure keeping C-element's state when the values of the inputs I_1, …, I_N do not match.

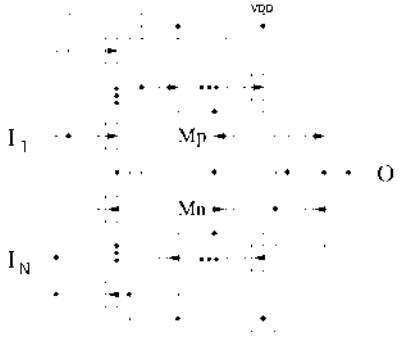


Fig. 2. Static CMOS implementation of the N-input C-element

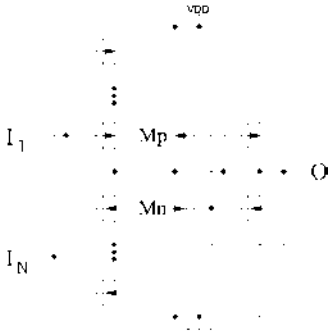


Fig. 3. Semi-static CMOS implementation of the N-input C-element

A semi-static C-element case uses the simplest two-transistor circuit providing C-element's memory feature. For the C-element to work correctly, the transistors Mn and Mp must be "weak": their channel width-to-length ratio should be less than that of the other transistors in the circuit. Let the channel widths of p-transistors ($W_{p,I}$) and n-transistors ($W_{n,I}$) match for all the same type transistors in the C-element input part, and their channel lengths ($L_{p,I}$, $L_{n,I}$) correspond to a CMOS process used. Then the memory part transistor sizes (W_{Mp}, W_{Mn}, L_{Mp}, L_{Mn}) should ensure the fulfillment of the relations:

$$\frac{L_{Mp}}{W_{Mp}} \geq N \cdot K_p \cdot \frac{L_{n,I}}{W_{n,I}}, \frac{L_{Mn}}{W_{Mn}} \geq N \cdot K_n \cdot \frac{L_{p,I}}{W_{p,I}}, \qquad (2)$$

where K_p and K_n are the coefficients depending on CMOS

process parameters. The output inverter transistor sizes are arbitrary. The disadvantage of the semi-static C-element implementation is a short circuit current through the transistors Mn and Mp when C-element switches. This short-circuit current is proportional to the width-to-length ratio of their channels.

The static C-element CMOS implementation is free of any short circuit current at C-element switches.

In the industrial standard cell libraries for 65nm and below CMOS process, the number of transistors connected in series in all cell schematics is less than four. Therefore, C-elements demonstrated in Fig. 2 and 3 cannot have more than three inputs (N ≤ 3). However, in [5], the multi-input C-element implementation was proposed, which has only two transistors connected in series and ensures minimal hardware complexity. Fig. 4 illustrates its CMOS circuit.

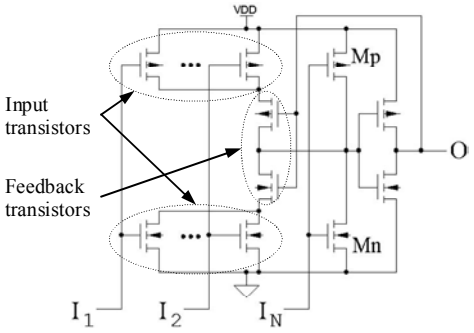


Fig. 4. Semi-static CMOS implementation of the multi-input C-element

For the multi-input C-element function correctly, the transistors Mn and Mp should also be "weak." They should not lead to a premature switching C-element at early changes of I_N when at least one of the other inputs remained in opposite to I_N state. Besides, the size of the transistors of the multi-input C-element should provide an acceptable "performance to short-circuit current value" ratio in the worst case.

The following transistor size ratios provide the necessary workability conditions for the multi-input C-element in typical 65-nm CMOS process [5]:

$$\begin{cases} \dfrac{L_{Mp}}{W_{Mp}} \geq K_{p,CM} \cdot \left(\dfrac{L_{n,in}}{W_{n,in}} + \dfrac{L_{n,FB}}{W_{n,FB}} \right), \\ \dfrac{L_{Mn}}{W_{Mn}} \geq K_{n,CM} \cdot \left(\dfrac{L_{p,in}}{W_{p,in}} + \dfrac{L_{p,FB}}{W_{p,FB}} \right). \end{cases} \quad (1)$$

Here $W_{p,in}$, $W_{n,in}$, $L_{p,in}$, $L_{n,in}$ are the width and length of p- and n-transistors gated by I_1, ..., I_{N-1} 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 C-element's state at time intervals when its inputs do not match; $K_{p,CM}$, $K_{n,CM}$ are coefficients depending on process-dependent parameters.

In all considered implementations, the two-transistor circuit drives output load.

III. C-ELEMENT IMPLEMENTATION IMMUNE TO SEUTS

C-element is a cell with memory. Therefore, it can remember the SEUT induced within it. In [6], authors have proposed a DICE-like C-element implementation for increasing its SEUT tolerance. Fig. 5 illustrates its circuit for the two-input semi-static C-element, while Fig. 6 shows the DICE-like static C-element circuit. The DICE-like C-element implementation prevents the storing single SEUT in the C-element and immunes C-element output O to any single SEUT at nodes N_2 and N_3.

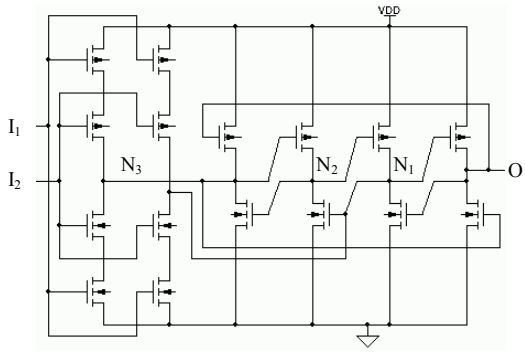


Fig. 5. DICE-like C-element CMOS semi-static implementation

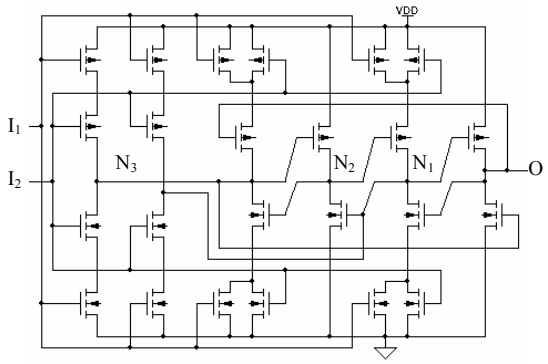


Fig. 6. DICE-like C-element CMOS static implementation

However, all DICE-like C-element implementations proposed in [6] do not provide full protection against SEUT appearing at node N_1. Fig. 7 illustrates the DICE-like semi-static C-element circuit that eliminates this drawback. It bases on four-transistor converters proposed in [7]. These converters drive node N_2 and output O, replacing pure inverters. Fig. 8 shows the analogous improved static C-element implementation. The same implementation also exists for multi-input C-element.

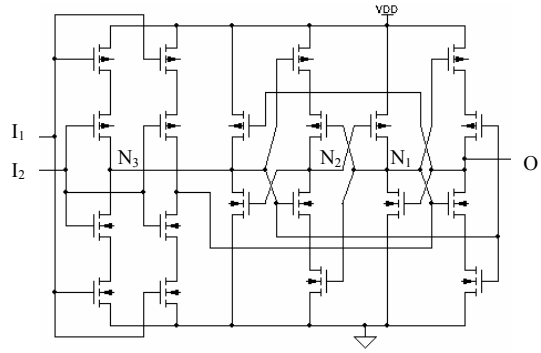


Fig. 7. Improved DICE-like CMOS semi-static C-element

Improved circuits provide immunity of the C-element output *O* to single SEUT at all internal nodes. Simulating C-element circuits in Spectre (Cadence) proves this fact. For this purpose, an ionization current pulse source emulates SEUT impacting one circuit node. The current pulse has the following parameters: 400-μA amplitude with any polarity, a 7-ps leading edge, a 200-ps drop, and a 200-ps "plateau" [3].

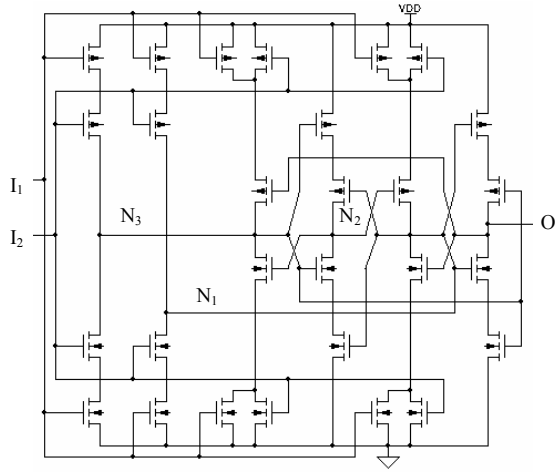


Fig. 8. Improved DICE-like CMOS static C-element

Fig. 9 illustrates responses of the compared C-element outputs to a single SEUT at nodes N_1. The labels, marking curves, reflect the figure's number showing the corresponding circuit. Fig. 10 demonstrates similar responses of the compared C-element outputs *O* to a single SEUT at nodes N_3.

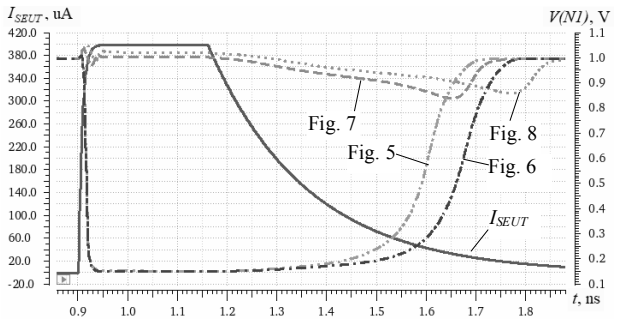


Fig. 9. DICE-like C-element output response to a single SEUT at N_1 node during high-level standby

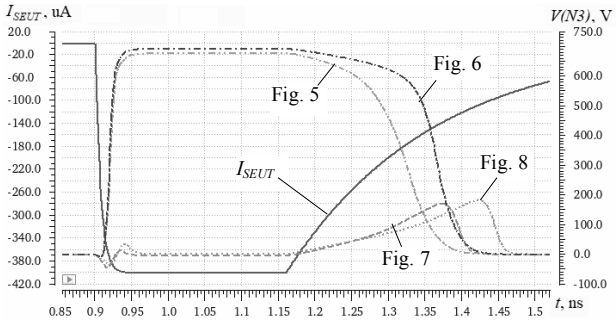


Fig. 10. DICE-like C-element output response to a single SEUT at N_3 node during low-level standby

One can see that C-element cases, shown in Fig. 7 and 8, with the four-transistor output converter, are entirely immune to SEUTs, unlike cases, shown in Fig. 5 and 6, with the two-transistor output converter. The last ones lose stored state for SEUT duration.

Thus, the implementation of the C-element output cascade by the four-transistor circuit guarantees masking SEUT induced at any internal node of the improved DICE-like C-element. However, SEUT can appear directly at output node *O*. To mask such SEUT case, C-element should have separate input pairs and two outputs corresponding to them.

Fig. 11 illustrates DICE-like semi-static C-element with the separate input pairs (I_1, I_2) and (J_1, J_2), and in-phase outputs O_1 and O_2. This C-element masks SEUT induced directly at any C-element output. In proper layout placement of the C-element symmetrical parts, a single SEUT affects only one C-element output (O_1 or O_2). Another C-element output keeps the correct level and prevents propagating SEUT through an indication subcircuit built on this C-element.

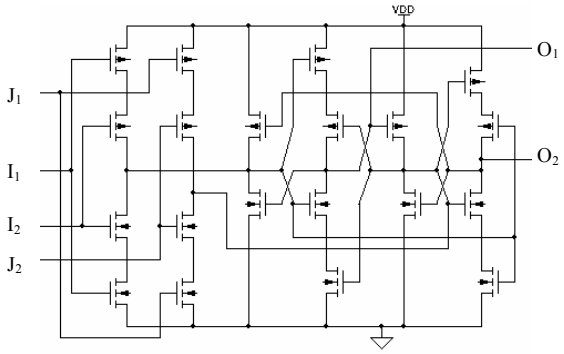


Fig. 11. DICE-like C-element with in-phase outputs O_1 and O_2

The input pairs I_k and J_k, k = 1, 2, are logically identical. They are generated by the in-phase outputs O_1 and O_2 of the previous C-elements in the indication subcircuit. Then a SEUT at one input (for example, I_1) is masked by its analog's (J_1) correct value.

Fig. 12 shows an example of an indication subcircuit built on the principles and circuits described above.

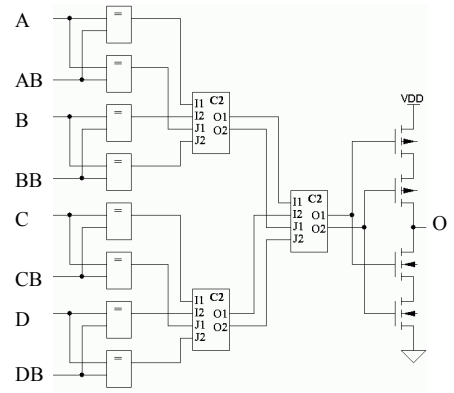


Fig. 12. SEUT protected indication subcircuit implementation

It indicates four dual-rail signals and bases on the XNOR cells in the first stage and C-elements with in-phase outputs (C2-cell, shown in Fig. 11) in the subsequent cascades.

Besides, proposed DICE-like improved C-elements have reduced consumption current (I_{CC}) during SEUT impact. Fig. 13 proves that current consumption pulse, caused by SEUT, in DICE-like C-elements offered in [6], has in several times larger amplitude than in the C-elements proposed in this paper.

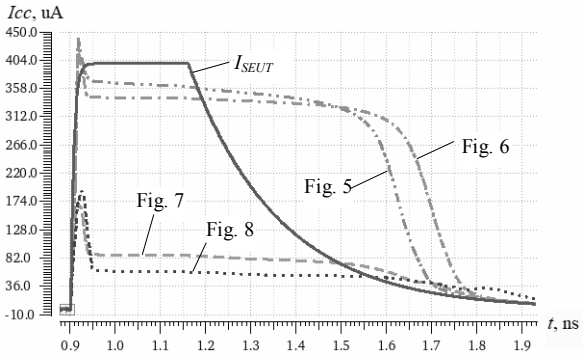


Fig. 13. Consumption current pulse in DICE-like C-elements at SEUT

Thus, indication subcircuit of any ST circuit becomes fully protected against single SEUTs due to using proposed design principles:

- XOR or XNOR cells on the first stage of the indication subcircuit,
- DICE-like C-elements with four-transistor output converter,
- in-phase inputs and outputs in each DICE-like C-element.

These principles implementation doubles indication subcircuit complexity. But it is the penalty for achieving full protection against SEUTs.

IV. Conclusion

The indication subcircuit largely determines ST circuit SEUT tolerance. Its complexity and layout area is up to half of the entire ST circuit complexity and area. Therefore, a SEUT affects an indication subcircuit with a probability close to that in the ST circuit rest part.

XOR or XNOR cells at the first stage of the indication subcircuit prevent the processing of the anti-spacer state as a working state of the dual-rail information signal.

The DICE-like C-element circuit improves the indication subcircuit immunity to single SEUTs. Using the four-transistor converter instead of the two-transistor one makes the DICE-like C-element entirely immune to single SEUTs at its internal and input nodes.

C-elements with in-phase inputs and outputs provide additional protection against single SEUT, which appears in C-element outputs. The four-transistor converter, used in the proposed DICE-like C-element as output driver, transforms in-phase signals into the unary indication signal if needed.

References

[1] Y.A. Stepchenkov, A.N. Kamenskih, Y.G. Diachenko, Y.V. Rogdestvenski, and D.Y. Diachenko, “Fault-Tolerance of Self-Timed Circuits,” in Proc.10th Int. Conf. on Dependable Systems, Services, and Technologies (DESSERT), Leeds, United Kingdom, 2019, pp. 41–44. Available: https://doi.org/10.1109/DESSERT.2019.8770047.

[2] D. Mavis and P. Eaton, “SEU and SET modeling and mitigation in deep submicron technologies,” inProc. IEEE Int. Reliability Physics Symp., April 15-19, 2007, Phoenix, Arizona, USA, pp. 293–305.

[3] D.E. Muller and W.S. Bartky, “A theory of asynchronous circuits,” in Proc. Int. Symposium on the Theory of Switching, Harvard University Press, April 1959, pp. 204–243.

[4] Y A. Stepchenkov, Y.G. Diachenko, Y.V. Rogdestvenski, Y.I. Shikunov, and D.Y. Diachenko, “Advanced Indication of the Self-Timed Circuits,” inProc. 2019 IEEE East-West Design & Test Symp. (EWDTS), Batumi, Georgia, September 13-16, 2019. pp. 207–212.

[5] I. A. Danilov, M. S. Gorbunov, A. I. Shnaider, A. O. Balbekov, Y. B. Rogatkin, and S. G. Bobkov, “DICE-based Muller C-elements for soft error tolerant asynchronous ICs,” inProc. 16th European Conf. on Radiation and Its Effects on Components and Systems (RADECS), 2016, Bremen, Germany, pp.1–4, Available:

[6] A. Eaton, “Single event upset immune logic family,” U.S. Patent 6 756 809, Jan. 29, 2004.

Ultragraph Model for ECE Component Partitioning

Elmar Kuliev
Computer-aided design department
Southern Federal University
Rostov-on-Don, Russia
elmar_2005@mail.ru

Dmitry Zaporozhets
Computer-aided design department
Southern Federal University
Rostov-on-Don, Russia
elpilasgsm@gmail.com

Daria Zaruba
Computer-aided design department
Southern Federal University
Rostov-on-Don, Russia
dvzaruba@sfedu.ru

Abstract — The article considers the modelling of a computational device at the design stage. One of the most labour-intensive problems is a partitioning problem which belongs to the class of NP-hard problems. In other words, there is no precise method for its addressing. The authors suggest an alternative way to model the device circuits as an ultragraph which simulates circuit components taking into account a direction of signal transmission. Thus, the suggested approach makes it possible to obtain an adequate model in terms of the correctness of information and its completeness. As an example, an ultragraph model of an amplifier is given both graphically and analytically. Thy ultragraph model is firstly adopted to ECE components partitioning problem. A problem statement is considered on the basis of the ultragraph model. A new encoding and decoding mechanism is developed to address the partitioning problem by a bioinspired algorithm. To confirm its effectiveness, a software is developed. The goal of the experiments is a calculation of CPU time and memory as well as of the comparison the ultragraph model with graph and hypergraph models. It is experimentally proved that the ultragraph model can reduce CPU time cost in comparison with other mathematical models.

Keywords—modelling, optimization, graph, ultragraph, bacterial foraging optimization.

I. Introduction

The design of computational devices is considered as one of the most labour-intensive stages on manufacture. At this stage, a schematic representation of elements in the device is interpreted as its geometric form.

The need for a high-quality elements partitioning is due to increasing requirements to the minituarization and speed of modern devices. The partitioning problem is to find a relative position of elements in the device in a way that all requirements have been met. This problem is NP-hard and belongs to the combinatorial and logic class of problems [1-3].

To formalize the designed object, a graph theory is useful and relevant to obtain the adequate model of the device in terms of completeness of information and its correctness.

II. Ultragraph as a Mathematical Model

A method of presenting a mathematical model plays a critical role in the development of computing devices and determines an adequacy of the model to the object, time and memory complexity, the convenience of the design, and its evaluability.

In this article it is proposed to use an ultragraph as a mathematical model for an adequate representation of the device.

The ultragraph is a graph $H(X, U, \Gamma_1, \Gamma_2)$ where incidence predicates $\Gamma_1(X, U)$ and $\Gamma_2(U, X)$ meet the following condition [4-8]:

$$\exists u_i \in U\left(\left|\Gamma_1 u_j\right| + \left|\Gamma_2 u_j\right|\right) > 2 \qquad (1)$$

Here, X is a set of vertices in the ultragraph H, U is a set of edges in the ultragraph H. In other words, the ultragraph contains at least one edge to which the total number of vertices to which it is incident, and which are incident to it is greater than two.

The ultragraph H with a set of vertices $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and a set of edges $U = \{u_1, u_2, u_3, u_4\}$ in Fig. 1.

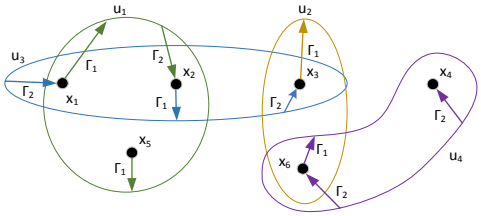


Fig. 1. Hybrid architecture

To illustrate the modelling process, it is proposed an amplifier circuit which is presented in Fig. 2 as an example [9,10].

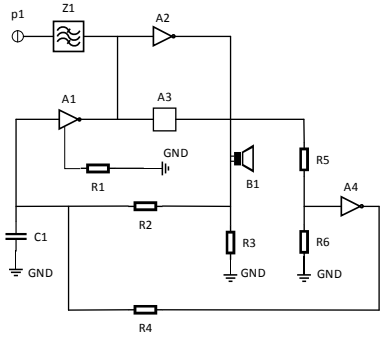


Fig. 2. The amplifier circuit

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

Let $X = \{x_1, \dots, x_{15}\}$ be a set of vertices that corresponds to components of the amplifier and $U = \{u_1, \dots, u_{13}\}$ is a set of edges which corresponds to its electrical circuit. The compliance between components of the amplifier and vertices of the ultragraph H is contained in Table 1.

TABLE 1. COMPLIANCE BETWEEN AMPLIFIER COMPONENTS AND VERTICES OF THE ULTRAGRAPH H

Element of the set X	Designation in the circuit	Description
x_1	p_1	source
x_2	Z_1	high-pass filter
x_3	A_2	low frequency power amplifier
x_4	A_1	reversing amplifier
x_5	A_3	negative feedback device
x_6	B_1	microphone
x_7	R_2	variable resistor
x_8	R_3	variable resistor
x_9	GND	ground
x_{10}	R_1	variable resistor
x_{11}	C_1	capacitor
x_{12}	R_4	variable resistor
x_{13}	A_4	reversing amplifier
x_{14}	R_5	variable resistor
x_{15}	R_6	variable resistor

The number of circuits in the amplifier needs to be defined to generate edges of the ultragraph. The amplifier includes 13 circuits: $u_1\{x_1, x_2\}$, $u_2\{x_2, x_3, x_4, x_5\}$, $u_3\{x_3, x_6\}$, $u_4\{x_6, x_7, x_8\}$, $u_5\{x_8, x_9\}$, $u_6\{x_4, x_{10}\}$, $u_7\{x_{10}, x_9\}$, $u_8\{x_{11}, x_9\}$, $u_9\{x_{11}, x_4, x_7, x_{12}\}$, $u_{10}\{x_{13}, x_{12}\}$, $u_{11}\{x_{13}, x_9, x_{14}, x_{15}\}$, $u_{12}\{x_{15}, x_9\}$, $u_{13}\{x_{14}, x_5\}$.

Incidence predicates Γ_1 and Γ_2 take on the value "true" on the following cases:

$$\Gamma_1(X,U) = (x_1,u_1),(x_2,u_2),(x_3,u_3),(x_4,u_6),(x_4,u_9),(x_5,u_{13}),(x_6,u_4),(x_7,u_9),(x_8,u_5),(x_{11},u_8),(x_{12},u_9),(x_{13},u_{10}),(x_{15},u_{12}).$$

$$\Gamma_2(U,X) = (u_1,x_2),(u_2,x_3),(u_2,x_4),(u_2,x_5),(u_3,x_6),(u_4,x_7),(u_4,x_8),(u_5,x_9),(u_6,x_{10}),(u_7,x_9),(u_8,x_9),(u_9,x_{11}),(u_{10},x_{12}),(u_{11},x_9),(u_{11},x_{14}),(u_{11},x_{15}),(u_{12},x_9),(u_{13},x_{14}).$$

The ultragraph which corresponds to the amplifier is shown in Fig. 3.

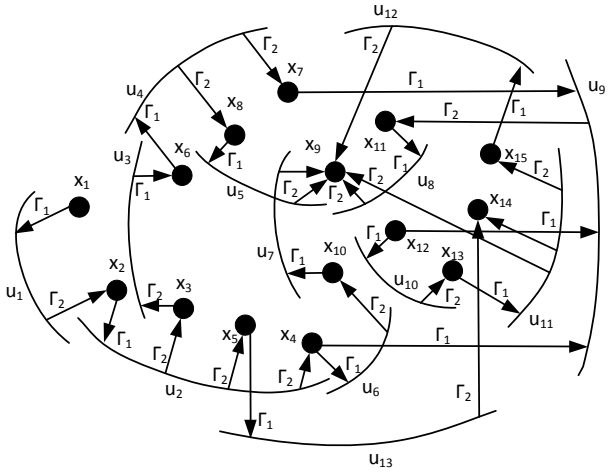


Fig. 3. The ultragraph which corresponds to the amplifier

Fig. 4 shows a bipartite graph, consistent with the ultragraph.

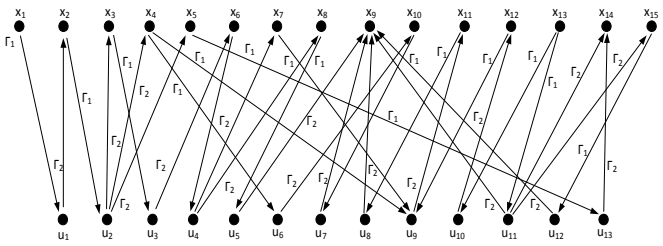


Fig. 4. The bipartate graph

Incident matrices A_1 and A_2 has the form:

$A_1 =$

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{13}
x_1	1	0	0	0	0	0	0	0	0	0	0	0	0
x_2	0	1	0	0	0	0	0	0	0	0	0	0	0
x_3	0	0	1	0	0	0	0	0	0	0	0	0	0
x_4	0	0	0	0	0	1	0	0	1	0	0	0	0
x_5	0	0	0	0	0	0	0	0	0	0	0	0	1
x_6	0	0	0	1	0	0	0	0	0	0	0	0	0
x_7	0	0	0	0	0	0	0	0	1	0	0	0	0
x_8	0	0	0	0	1	0	0	0	0	0	0	0	0
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0	1	0	0	0	0	0	0
x_{11}	0	0	0	0	0	0	0	1	0	0	0	0	0
x_{12}	0	0	0	0	0	0	0	0	1	0	0	0	0
x_{13}	0	0	0	0	0	0	0	0	0	1	0	0	0
x_{14}	0	0	0	0	0	0	0	0	0	0	0	0	0
x_{15}	0	0	0	0	0	0	0	0	0	0	0	1	0

$A_2 =$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
u_1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
u_2	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
u_3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
u_4	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
u_5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
u_6	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
u_7	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
u_8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
u_9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
u_1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
u_1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
u_1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
u_1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

For illustrative purposes and the convenience of evaluations, matrices A_1 and A_2 come down to a generalized matrix R under the following formula:

$$R = A_1^T + (-1) * A_2. \quad (2)$$

Then, the generalized matrix R has a form

$$R = \begin{array}{c|cccccccccccccc} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ u_1 & 1 & -_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ u_2 & 0 & 1 & -_1 & -_1 & -_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ u_3 & 0 & 0 & 1 & 0 & 0 & -_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ u_4 & 0 & 0 & 0 & 0 & 0 & 1 & -_1 & -_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ u_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -_1 & 0 & 0 & 0 & 0 & 0 \\ u_6 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -_1 & 0 & 0 & 0 & 0 \\ u_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -_1 & 1 & 0 & 0 & 0 & 0 \\ u_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -_1 & 0 & 1 & 0 & 0 & 0 \\ u_9 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & -_1 & 1 & 0 & 0 \\ u_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -_1 & 1 & 0 \\ u_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ u_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -_1 & 0 & 0 & 0 & 0 & 1 \\ u_1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -_1 & 0 \end{array}$$

Furthermore, the ultragraph model can accurately and adequately assesses the number of interconnections as well as contains complete information on the circuit.

III. Partitioning Problem Statement

In the article, the components partitioning in electronic devices is considered in terms of a system approach and in the light of the progress in new information and telecommunication technologies. The partitioning is a grouping of low-level elements to obtain high-level elements taking into account the criteria [11, 12].

The components partitioning is considered as a decomposition of the ultragraph $H(X, U, \Gamma_1, \Gamma_2)$ into parts $H_i = (X_i, U_i, \Gamma_{1i}, \Gamma_{2i}), X_i \subseteq X, U_i \subseteq U, i \in I = \{1,2, \dots, l\}$, where l is the number of parts to which the ultragraph H is split.

In other words, a set of parts $P(H) = \{H_1, H_2, \dots, H_i, \dots, H_l\}$ is the ultragraph H partitioning if any part is not empty and, for every pair of parts from $P(H)$, the intersection of vertex sets is an empty set, but the intersection of edge sets is not empty, as well as the union of all parts l is equal to the ultragraph H.

Let each subset H_i contains n elements $H_i = \{h_1, h_2, \dots, h_n\}$, where $n = |X|$. Then, the ultragraph H partitioning comes down to the partitioning $H_i \in H$ meeting the following conditions and limits:

$$(\forall H_i \in H)(H_i \neq \varnothing),$$

$$\left(\begin{array}{c}(\forall H_i, H_j \in H) \\ \left[H_i \neq H_j \rightarrow X_i \cap X_j = \varnothing\right] \wedge \\ \wedge \left[(U_i \cap U_j = U_{ij}) \vee (U_i \cap U_j = \varnothing)\right]\end{array}\right),$$

$$\bigcup_{i=1}^{S} H_i = H, \bigcup_{i=1}^{m} U_i = U, \bigcup_{i=1}^{n} X_i = X.$$

Here, $m = |U|$.

As an optimization criterion, it is used the number of interconnections between subsets H_i and H_j. In this article, optimization criterion is defined as follows:

$$K = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}, (i \neq j). \quad (3)$$

Here, c_{ij} is the number of interconnections between subsets H_i and H_j, K is a total number of interconnections between all subsets in the ultragraph H [13-16].

The goal of optimization is to minimize the K ($K \rightarrow \min$).

It should be noted that the partitioning problem belongs to the combinatorial and logic class of problems, i.e., a search for optimal solutions is related to a large number of possible partitioning. Therefore, the components partitioning with predetermined conditions and limits is a mechanism to prepare the model for further design stages.

IV. New Encoding and Decoding Mechanism

Encoding and decoding of alternative solutions is a key issue in meeting science and technology challenges. A unified approach to data representation for bioinspired partitioning algorithm is suggested to speed up encoding and decoding of alternative solutions and calculate the objective function (OF). Each partition is encoded as a sequence of blocks. The alternative solution (a chromosome) is represented as a sequence of block numbers, divided into groups [17, 18].

Suppose, the circuit should be divided into two blocks.

As we know, a generation of an initial population is required to initialize the bioinspired algorithm. In this case, the initial population is shown in Fig. 5.

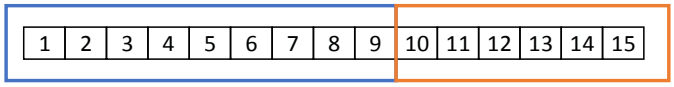


Fig. 5. Encoded initial partition

The first block contains $x_1 - x_9$ vertices, the second block - $x_{10} - x_{15}$. The bipartite graph illustrates the initial partition and objective function calculation as shown in Figure 6.

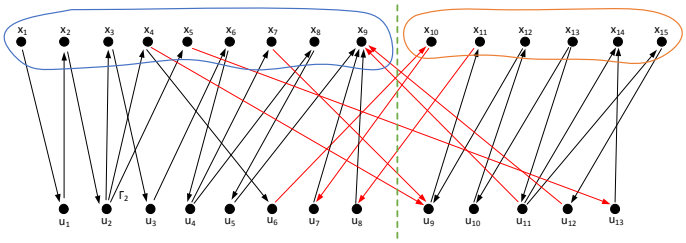


Fig. 6. The initial bipartite graph

As shown in Fig. 6, the number of interconnections is equal to 8 i.e., the OF=8.

The bioinspired partitioning algorithm provides the alternative solution presented in Fig. 7.

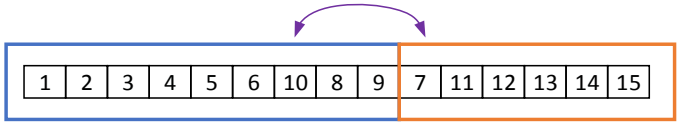


Fig. 7. Encoded solution after the bioinspired partitioning algorithm

As a result, the number of interconnections has decreased, i.e. OF=6, as shown in Fig. 8.

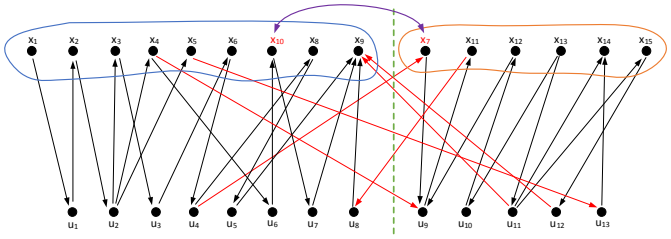


Fig. 8. The bipartite graph after the bioinspired partitioning algorithm

It should be noted, CPU time cost can be reduced using the developed encoding and decoding mechanism.

V. EXPERIMENTS

A classical bacterial foraging optimization algorithm [19-21] focused on the partitioning problem was implemented for pilot studies. The algorithm was realized as a C++ application which can implement three mathematical models of the circuit, namely, ultragraph, hypergraph, and graph models. Experiments were carried out to estimate memory and CPU time costs due to extraction and processing of information. It can be an objective function of calculation and the encoding and decoding process. Note that the difference between graph and ultragraph models is to use the number of the element with a negative sign as a source element. This fact places a limit on a maximum size of input data, since the most significant bit becomes a sign bit.

To estimate total memory cost, there is considered memory to store edges between vertices and information about whether a vertex in a particular block (a set B).

For the graph model, it is used an adjacency matrix which stores data about adjacency of any two vertices but not information about edges. The adjacency matrix is represented as $R^1_{N\times N}$, where N is a number of blocks. A redundancy of this model is equal to

$$RETUND^1 = N * (N-2), \tag{4}$$

In hypergraph and ultragraph models, an edge can be incident to two or more vertices. Then, the adjacency matrix is represented as $R^2_{M\times N}$, where N and M is a number of elements and edges, correspondingly. The redundancy of these models depends on N, M, and an average number of elements that incident to an edge. In this case, there is an opportunity to avoid the redundancy due to storage of an edge as a vector pointed to a circuit element.

$$RETUND^2 = 0. \tag{5}$$

To determine the computational complexity of objective function calculation, let assume that a set of vertices X is divided into a set of blocks $B\{b_i\}$, where for any $x_i \in X$ there is a $b_i \in B$, $|B| > 2$. B is a set as an incidence matrix of elements X to blocks B.

For a graph model, the objective function is calculated as follows:

1. Let $x_i \in X$ (a constant).
2. To find b_i to which x_i belongs to. ($|B|$ operations)
3. To find the vertex x_i' adjacent to the x_i. ($|X|$ operations)
4. To find the block b_j' to which the vertex x_i' belongs to. ($|B|$ operations)
5. Check whether b_j and b_j' are equal. If they are not equal, then a block interconnection counter is increased of 1. (a constant)

Furthermore, the computational complexity of the algorithm can be calculated as follows:

$$O(N,B) = (2B + X) * X \tag{6}$$

The number of blocks B is usually much less than a number of elements X, and it can be ignored. Then, the computational complexity of objective function calculation is quadratic and is represented as $O(n^2)$.

In terms of hypergraph and ultragraph models, to calculate the number of block interconnections, it is determined the number of blocks b_i to which an edge $e_i \in E$ belongs to.

In this case, the objective function is calculated as follows:

1. Let $e_i \in E$ (a constant).
2. For each x_{ij} from e_i is found a block b_k to which x_{ij} belongs to ($|B| * |e_i|$ operations)
3. To calculate the number of unique blocks (constant)

Since $x_i \in X$ can belong to several edges simultaneously, then $|B| * |e_i|$ operations are calculated in the following way:

$$|B| * |e_i| * |p|, \tag{7}$$

where p is an average value of a repetition of the element in several edges.

Therefore, for hypergraph and ultragraph models, the computational complexity of the algorithm is

$$O(N,E,B) = B * E. \tag{8}$$

Here, the number of blocks B can also be ignored, and the computational complexity is linear and is represented as $O(n)$.

Fig. 9 and Table 2 shows a time dependency for graph, hypergraph and ultragraph models. For each test circuit, it was generated 1000 solutions in randomly way.

TABLE 2. CPU TIME DEPENDANCE OF THE SOLUTIONS ON THE NUMBER OF ELEMENTS IN THE CIRCUIT

S. No.	Number of elements	CPU time, s	
		Graph model	Hypergraph
1	12 752	0,09	0,04
2	19 601	0,21	0,07
3	23 136	0,30	0,08
4	27 507	0,42	0,09
5	29 347	0,48	0,10
6	32 498	0,59	0,11
7	45 926	1,17	0,15
8	51 309	1,46	0,17
9	53 395	1,58	0,18
10	69 429	2,68	0,23
11	70 558	2,77	0,24
12	71 076	2,81	0,24

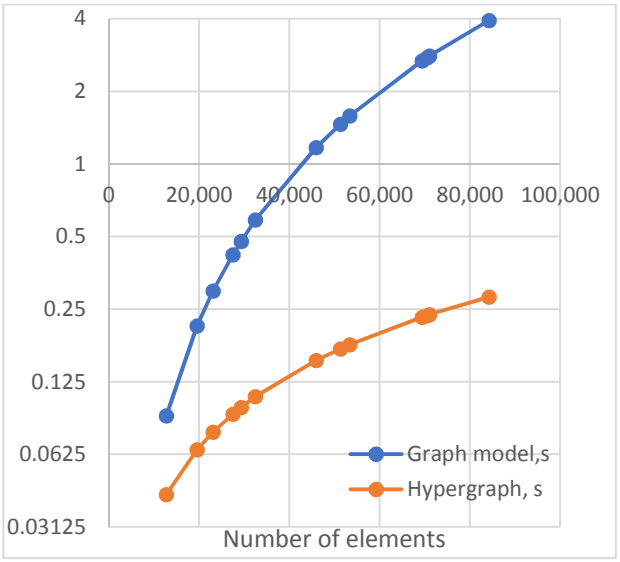


Fig. 9. CPU Time dependency

VI. CONCLUSION

The ultragraph model has been applied for the first time to address the ECE components partitioning problem. The use of the ultragraph model is an alternative way for circuit representation which enable to accurately estimate the number of connections in the circuit taking into account a signal direction. The ultragraph fully meets the requirements for a simulated object as well as contributes to the formalization and algoritmization the circuit model. The encoding and decoding mechanism has been developed to the partitioning problem. To carry out computational experiments, it has been used in a classical bacterial partitioning algorithm. On the basis of conducted theoretical and experimental investigations, the following conclusions can be made:

1. In terms of design problems, graph models have low efficiency, since a memory redundancy for a model storage is quadratic. The time complexity of the objective function calculation is also quadratic and can be represented as $O(n^2)$.

2. For hypergraph and ultragraph models, time complexity is linear, and memory redundancy is equal to 0.

3. Unlike the hypergraph model, the ultragraph considers a signal direction in a circuit using negative numbers to store it. In this case, a maximum number of possible input data is reduced twice, since a sign bit is used for a sign store.

ACKNOWLEDGMENT

This research is supported by a grant of the Russian Foundation for Basic Research (RFBR), the project # 19- 01- 00059.

REFERENCES

[1] J. Kacprzyk, V.M. Kureichik, S.P. Malioukov, V.V. Kureichik and A.S. Malioukov, "General questions of automated design and engineering," Studies in Computational Intelligence, vol. 212, 2009, pp. 1-22.

[2] C.J. Alpert, D.P. Mehta and S.S. Sapatnekar, Handbook of Algorithms for Physical Design Automation, 2009

[3] I. L. Markov, J. Hu, M.-C. Kim, Progress and Challenges in VLSI Placement Research. In: Proceedings of the IEEE, vol. 103, pp. 1985-2003 (2015)

[4] D. Goncalves, B.B.Uggioni, "Ultragraph shift spaces and chaos", in Bulletin des Sciences Mathematiques, vol. 158, 2016, #102807.

[5] G.G. de Castro, D. Goncalves, "KMS and Ground States on Ultragraph C*-Algebras", in J. Integral Equations and Operator Theory, 2018, art. no. 63.

[6] T. Bates, T.M. Carlsen, D. Pask, "C∗-algebras of labelled graphs III-K-theory computations," in Ergodic Theory and Dynamical Systems, vol. 37 (2), 2017, pp. 337-368.

[7] V.A. Ovchinnikov, The mathematical models of objects for structural creation tasks, Science and Education of the Bauman MSTU, 2009, no. 3.

[8] T. Katsura, P.S. Muhly, A. Sims, M.Tomforde, "Graph algebras, Exel-Laca algebras, and ultragraph algebras coincide up to Morita equivalence" in Journal fur die Reine und Angewandte Mathematik, vol. 640, 2012, pp.135-165.

[9] B.M. Ballweber, R.Gupta, D.J. Allstot, "A fully integrated 0.5-5.5-GHz CMOS distributed amplifier" in IEEE Journal of Solid-State Circuits, vol. 35 (2), 2000, pp. 231-239.

[10] A. Richter, S. Dris, I. Koltchanov, S. Alreesh, D. Yevseyenko, E. Sokolov, "Optical interconnects for datacenter links: Design and modeling challenges", in Proceedings of SPIE - The International Society for Optical Engineering, 11286, 2020, № 1128617,

[11] V.V. Kureichik, V.V. Kureichik and D.V. Zaruba, "Hybrid approach for graph partitioning" Advances in Intelligent Systems and Computing, vol. 573, 2017, pp. 64-73

[12] V. Kureichik, V. Bova, V.Kureichik, "Hybryd Approach for Computer-Aided Design Problems" in 2019 International Seminar on Electron Devices Design and Production, SED 2019 – Proceedings, 2019, #8798406.

[13] V.V. Kureichik, V.V. Kureichik, D.V. Zaruba, "Partitioning of ECE schemes components based on modified graph coloring algorithm" in Proceedings of IEEE East-West Design and Test Symposium, EWDTS 2014, № 7027062,

[14] V.M. Kureichik, V.V. Kureichik, "A genetic algorithm for graph partitioning" in J. of Computer and Systems Sciences International, 38 (4), pp. 580-588.

[15] E. Kuliev, V. Kureichik, V. Kureichik Jr. "Mechanisms of swarm intelligence and evolutionary adaptation for solving PCB design tasks" in 2019 International Seminar on Electron Devices Design and Production, SED 2019 - Proceedings, № 8798449.

[16] D. Zaruba, D. Zaporozhets, E.Kuliev, "Parametric optimization based on bacterial foraging optimization", in J. Advances in Intelligent Systems and Computing, vol. 573, 2017, pp. 54-63.

[17] V.Kureichik, D. Zaporozhets, D. Zaruba, "Generation of bioinspired search procedures for optimization problems", in Application of Information and Communication Technologies, AICT 2016 - Conference Proceedings, № 7991822, .

[18] L.A. Gladkov, V.V. Kureichik,. Y.A. Kravchenko, "Evolutionary algorithm for extremal subsets comprehension in graphs" in World Applied Sciences Journal, vol. 27 (9), 2013, pp. 1212-1217.

[19] Zhao, W., Wang, L. An effective bacterial foraging optimizer for global optimization (2016) Information Sciences, 329, pp. 719-735.

[20] Liu, Y., Passino, K.M. Biomimicry of social foraging bacteria for distributed optimization: Models, principles, and emergent behaviors (2002) Journal of Optimization Theory and Applications, 115 (3), pp. 603-628.

[21] Karpenko A.P. Modern algorithms of search optimization. Algorithms inspired by nature. Moskow, Russia. 2014.P. 446.

The Method of Increasing of CMRR for CJFET Dual Differential Input Stages for the Tasks of Processing Sensor Signals Under Conditions of Cryogenic Temperatures and Penetrating Radiation

Nikolay Prokopenko
Don State Technical University, Institute for Design Problems in Microelectronics of RAS, Rostov-on-Don, Zelenograd, Russia
prokopenko@sssu.ru

Alexey Zhuk
Don State Technical University Rostov-on-Don, Russia
alexey.zhuk96@mail.ru

Ilya Pakhomov
Don State Technical University Rostov-on-Don, Russia
ilyavpakhomov@gmail.com

Petr Budyakov
JSC "S&PE "PULSAR" Moscow, Russia
budyakovp@gmail.com

Alexey Titov
Southern Federal University Rostov-on-Don, Russia
alex.evgeny.titov@gmail.com

***Abstract*—The architecture of the CJFET dual differential input stage (DDS) architecture is available for converting differential output sensor signals. DDS is proposed, in which the circuitry solutions are available. They provide the increase of the CMRR, including in severe application (low temperatures, radiation). It is shown that due to the special construction of the output current adder in the considered DDS, the transconductance of the input common-mode signal is significantly reduced to the output DSS. The versions of practical circuits of CJFET dual differential input stages based on current mirrors and "folded" cascodes with the increased CMRR and the results of their computer simulation at cryogenic temperatures in LTspice simulation software on CJFET models are presented. The proposed DDSs are recommended to be used in the structure of low-noise analog interfaces of sensors of various physical quantities applied in medicine, high-energy physics, and space instrument engineering.**

Keywords—analog sensor interfaces, cryogenic electronics, current mirrors, dual differential input stage, "folded" cascodes, input common-mode rejection ratio, junction field-effect transistors, penetrating radiation

I. Introduction

The common-mode noise immunity of the classical dual differential input stages (DDS) and operational amplifiers (OA) based on them has a significant effect on the precision of differential (DOA), differential difference (DDOA) and instrumentation (IA) amplifiers [1-17]. The solution to the problem of increasing the CMRR is even more complicated when working with analog interfaces that use DDS (OA, DDOA, IA) at low temperatures [17-20] and exposure to the penetrating radiation [21-26].

In this regard, the search for new DDS architectures with the increased input common-mode rejection, including in severe application, is of current interest.

The purpose and novelty of the paper is to research the limit parameters of the new structure of the CJFET dual differential input stage, in which due to the organization of special compensating channels in the output current adder, the errors from the input common-mode signal are reduced and, as a result, the CMRR is increased. Moreover, the use of the Si and GaAs [27] JFET active elements enables to create analog devices for operation at cryogenic temperatures [17-20] and exposure to the neutron flux [21-26].

The study has been carried out at the expense of the grant from the Russian Science Foundation (Project No. 16-19-00122-P).

II. Functional Scheme of DDS with the Increased Input Common-Mode Rejection

The feature of the proposed DDS [28-29] circuit shown in Fig. 1, consists in creating (for a common-mode signal) of $v_c=v_{c1}=v_{c2}$ parallel channels for transmitting output current increments through current-stabilizing one-port devices (R1, M3 and R2, M6) to the output DDS Out.Σ_1 with correctly selected current ratios K_1, K_2, K_3 and K_4, which are provided by the circuitry of adder Σ_1.

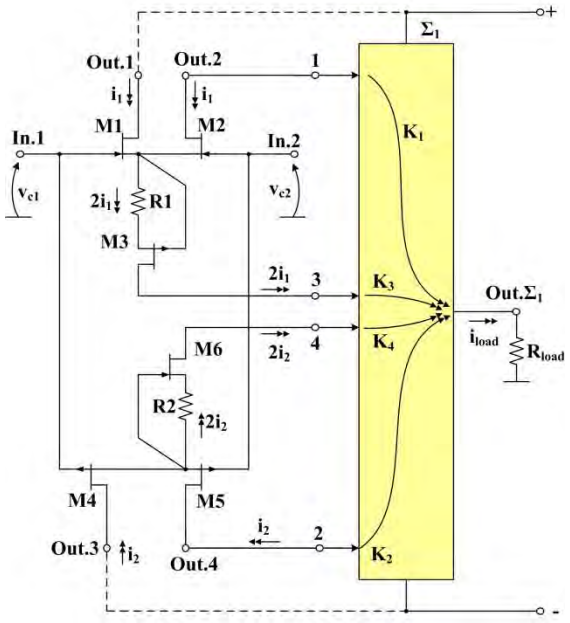


Fig. 1. Functional scheme of DDS with the increased CMRR.

On the base of the first Kirchhoff's law, it can be found that the output current of adder Σ_1 in the scheme of Fig. 1,

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

due to the change in the input common-mode signal $v_c=0.5(v_{c1}+v_{c2})$ will be zero, if:

$$(2K_3\text{-}K_1)g_{cm.1} = (K_2\text{-}2K_4)g_{cm.2}, \quad (1)$$

where $g_{cm.1}=i_1/v_c$, $g_{cm.2}=i_2/v_c$ are transconductances of the v_c to DDS Out.2 and Out.4.

In the general case, $g_{cm.1}\neq g_{cm.2}$. Therefore, to reduce the influence of the current-stabilizing one-port devices consisting of elements R1 and M3, as well as R2 and M6, on the transconductance v_c to the output node Out.Σ_1, it is necessary to provide the following current ratios using the circuitry of adder Σ_1:

$$\begin{cases} K_3 = 0.5K_1 \\ K_4 = 0.5K_2 \end{cases}. \quad (2)$$

In this case, K_1 and K_3, as well as K_4 and K_2 may change or may not change the phase of the current signal. In the first case, adder Σ_1 is implemented on the basis of inverting current mirrors for which the current transmission coefficient $K_1=K_3=-1$. In the second case, the subcircuit of adder Σ_1 must contain the noninverting current amplifiers (CA) with $K_1=K_3=+1$. In practice, such CAs are implemented in the basis of the CJFET "folded" cascodes [30-32].

In the circuit of Fig. 1, the transmission of the differential signal $v_d=v_{c1}-v_{c2}$ to the Out.Σ_1 through coefficients K_3 and K_4 of adder Σ_1 can be neglected, since the equivalent resistance of the current-stabilizing one-port devices on elements M3 and R1, as well as M6 and R2, significantly exceeds the resistance of the sources of input transistors M1-M4. However, for significant amplification of the differential signal ($v_d=v_{c1}-v_{c2}$), current ratios K_1 and K_2 in adder Σ_1 must equally change the phase of the input current signal, i.e., be either inverting or noninverting:

$$\text{sign}K_1=\text{sign}K_2, \quad (3)$$

where signK is a sign function, which describe DDS's output current signal phase change when its transmission through adder Σ_1. Provided that $\text{sign}K_i=-1$, if phase is changing and $\text{sign}K_i=+1$, if phase isn't changing.

The last equation is an additional restriction on the choice of K_1 and K_2.

Similarly, you can impose limitations on the signs of current ratios K_3, K_4 for the corresponding inputs 3, 4 of adder Σ_1:

$$\text{sign}K_3=\text{sign}K_4. \quad (4)$$

Thus, the synthesis of practical analog circuits with the architecture of Fig. 1 is reduced to providing the required numerical values of modules K_1, K_2, K_3, K_4, as well as their signs, which provide the increased CMRR:

$$CMRR^{-1} = \frac{g_{cm}}{g_d}, \quad (5)$$

where g_{cm} – input common-mode signal transconductance DDS to the output Out.Σ_1; g_d –input differential signal transconductance DDS to the output Out.Σ_1, besides:

$$g_{cm}=g_{cm.1}(K_1\text{-}2K_3) + g_{cm.2}(K_2\text{-}2K_4), \quad (6)$$

$$g_d \approx \frac{S_1S_2}{S_1+S_2}K_1 + \frac{S_4S_5}{S_4+S_5}K_2, \quad (7)$$

S_i - slope of the drain-gate characteristic of the i-th FET.

After transform of equations (5)–(7) for the case, when $K_1\approx K_2\approx 1$, we can obtain:

$$CMRR^{-1} \approx \frac{g_{cm.1}(K_1-2K_3)+g_{cm.2}(K_2-2K_4)}{S_{12}+S_{34}}, \quad (8)$$

where $S_{12} \approx 0.5S_1 \approx 0.5S_2$, $S_{34} \approx 0.5S_3 \approx 0.5S_4$.

Thus, the gain on the CMRR, which gives the circuit of Fig. 1, in comparison $CMRR_c$, when $K_3=0$, $K_4=0$:

$$N_R = \frac{CMRR}{CMRR_c} \approx \frac{g_{cm.1}K_1+g_{cm.2}K_2}{g_{cm.1}(K_1-2K_3)+g_{cm.2}(K_2-2K_4)}. \quad (9)$$

If $g_{cm.1}\approx g_{cm.2}\approx g_{cm}$, $K_1\approx 1$, $K_2\approx 1$, then at fulfillment of conditions (2) the CMRR in practical schemes Fig.1 reaches one or two orders:

$$N_R = \frac{2}{K_1-2K_3+K_2-2K_4} \gg 1. \quad (10)$$

III. Dual Input Stage Based on the "Folded" Cascodes

In the DDS of this subclass (Fig. 2), the required transfer ratios ($K_3=K_4=0.5$), satisfying equation (2), are realized by dividing the drain current of transistors M3 and M6 by two, as well as using the "folded" cascades on transistors M7, M8.

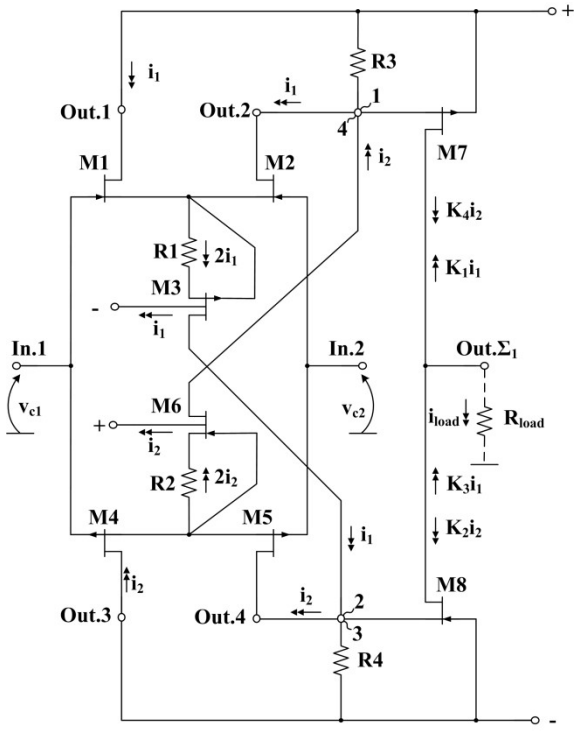


Fig. 2. The CJFET DDS with the increased CMRR oriented on the "folded" cascode.

The results of computer simulation (Fig.3) of common–mode signal transconductances of the circuit of

Fig. 2 in the frequency range are shown, that the limiting values of CMRR depend on the symmetry of the static mode with respect to the gate-drain voltage of the transistors of differential pairs M1–M2, M4–M5, as well as the non–identity of the static voltages at the drains of the composite transistors M3 and M6.

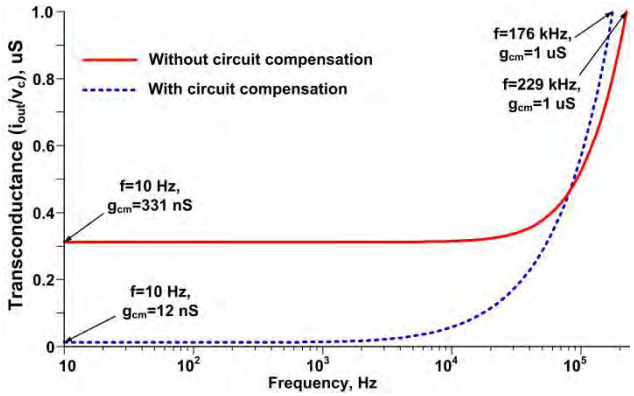


Fig. 3. The frequency dependence of the transconductance g_{cm} of the scheme of Fig. 2 with compensation and without circuit compensation.

In the circuit of Fig. 4, the required values of transfer ratios K_3=K_4=0.5 are ensured due to the use of two–channel "folded" cascodes on transistors M7, M9 and M8, M10.

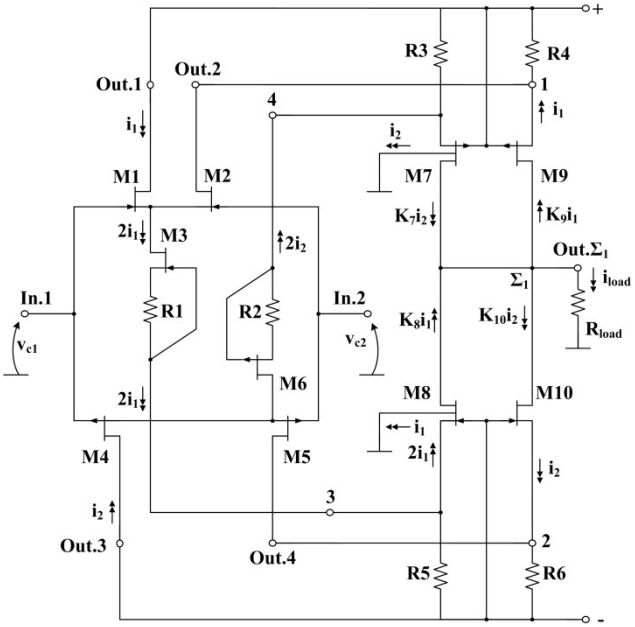


Fig. 4. The CJFET DDS based on two–channel "folded" cascodes.

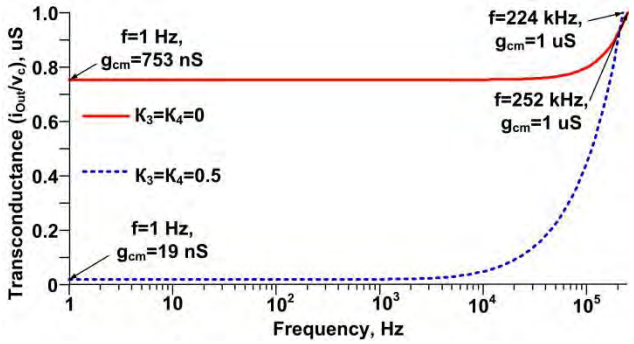


Fig. 5. The frequency dependence of the conductivity transmission g_{cm} of the scheme of Fig. 4 with common mode compensation (K_3=K_4=0.5) and without compensation, when K_3=K_4=0.

IV. Dual Input Stage Based on Current Mirrors

If in adder Σ_1 we use current mirrors CM1, CM2 with two inputs, through which current transfer A_1=K_1=K_2=-1 and A_2=K_3=K_4=-0.5 is provided, then the CMRR increase is realized in the circuit of Fig. 6.

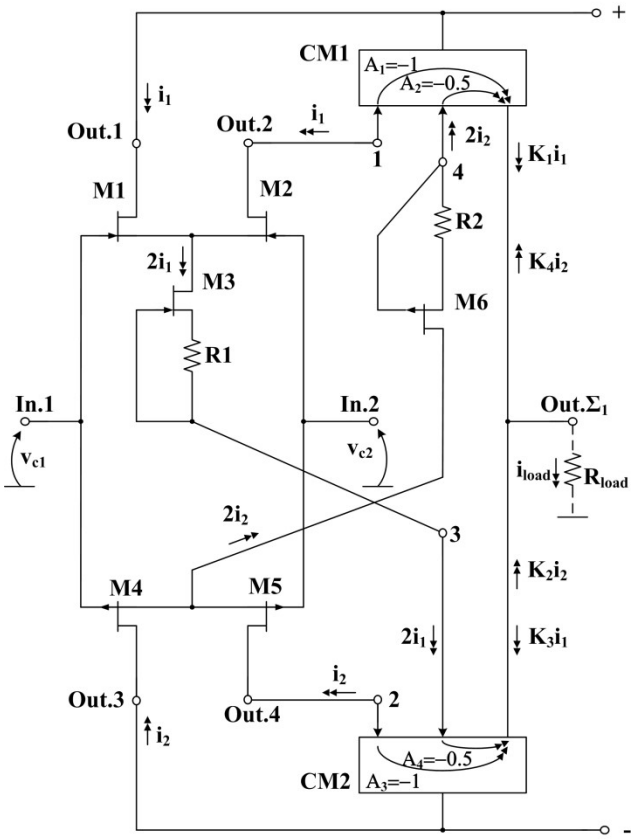


Fig. 6. The CJFET DDS based on Current Mirrors CM1, CM2.

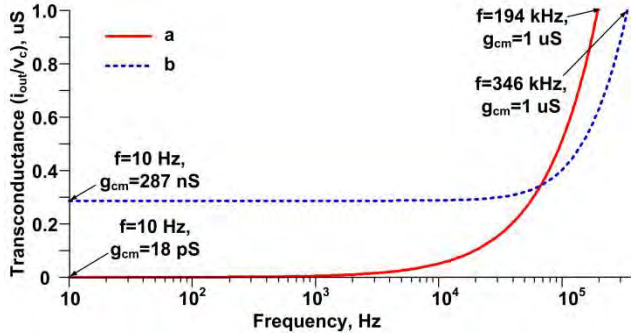


Fig. 7. The frequency dependence of the transconductance g_{cm} of the scheme of Fig. 6 at K_2=-0.5, K_4=-0.5 (a) and K_2=0, K_4=0 (b).

The specificity of the scheme of Fig.8 is that here required transfer ratios K_3 and K_4 of adder Σ_1 are provided by dividing in half the drain currents of transistors M3 and M6.

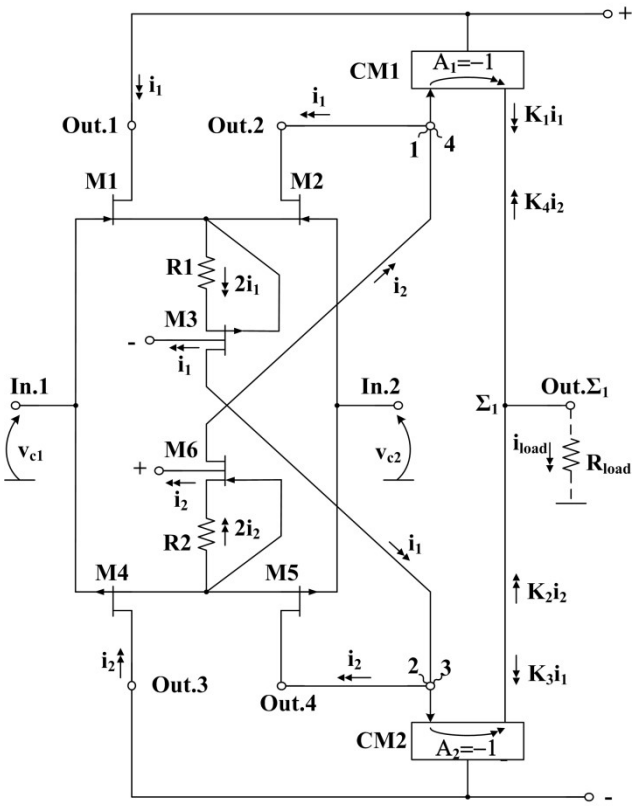


Fig. 8. The CJFET differential stage with the increased CMRR .

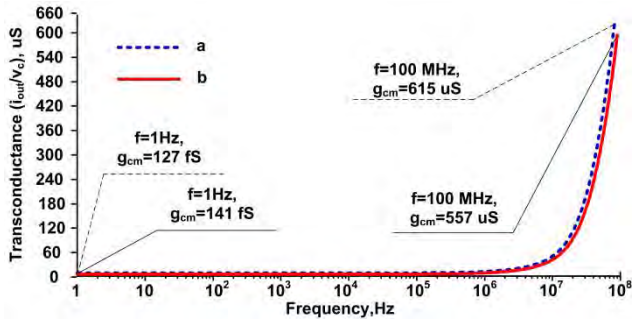


Fig. 9. The frequency dependence of the conductivity transmission g_{cm} of the input common–mode signal to the DDS output Fig. 8 at t=27°C (a) and -197°C (b).

V. Conclusion

A new architecture CJFET for the dual differential input stage has been developed. To compensate the errors from the input common–mode signal connected with the influence of the output resistances of the reference current sources a specified channel is provided, which ensures the stabilization of the static mode of the DDS transistors.

The computer simulation results show that the considered options for constructing a dual differential input stage based on inverted cascodes and current mirrors reduce by 2–3 orders of magnitude the error from the input common–mode signal, brought to the input of the amplifier.

The considered DDS circuits are recommended for encoders with differential output.

References

[1] Shruti Jain, “To Design High CMRR, High Slew rate Instrumentation Amplifier using OTA and CDTA for Biomedical Application”, *International Journal of Engineering Research*, Volume No.2, Issue No. 5, pp : 332–336, 01 Sept. 2013

[2] Eminoglu Selim, Petersen Anders K., “Low-voltage differential signaling receiver with common mode noise suppression”; US Patent Application Publication 2010/0013537; Jan. 21, 2010

[3] Botker Thomas Lloyd, “Common mode rejection ratio versus frequency in instrumentation amplifier”, US Patent 8.742.848, Jun. 3, 2014

[4] R. L. Schoenfeld, "Common-Mode Rejection Ratio-Two Definitions," *in IEEE Transactions on Biomedical Engineering*, vol. BME-17, no. 1, pp. 73-74, Jan. 1970. doi: 10.1109/TBME.1970.4502691

[5] M. Konar, R. Sahu and S. Kundu, "Improvement of the Gain Accuracy of the Instrumentation Amplifier Using a Very High Gain Operational Amplifier," *2019 Devices for Integrated Circuit (DevIC)*, Kalyani, India, 2019, pp. 408–412. doi: 10.1109/DEVIC.2019.8783414

[6] J. Oreggioni, A. A. Caputi and F. Silveira, "Current-Efficient Preamplifier Architecture for CMRR Sensitive Neural Recording Applications," *in IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 3, pp. 689–699, June 2018. doi: 10.1109/TBCAS.2018.2826720

[7] F. Centurelli, P. Monsurrò, G. Parisi, P. Tommasino and A. Trifiletti, "A Topology of Fully Differential Class-AB Symmetrical OTA With Improved CMRR," *in IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 11, pp. 1504–1508, Nov. 2018. doi: 10.1109/TCSII.2017.2742240

[8] E. H. T. Shad, M. Molinas and T. Ytterdal, "Modified Current-reuse OTA to Achieve High CMRR by utilizing Cross-coupled Load," *2019 15th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)*, Lausanne, Switzerland, 2019, pp. 13–16. doi: 10.1109/PRIME.2019.8787797

[9] J. Oreggioni, P. Castro-Lisboa and F. Silveira, "Enhanced ICMR amplifier for high CMRR biopotential recordings," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 3746–3749. doi: 10.1109/EMBC.2019.8856656

[10] S. Lee et al., "A 110dB-CMRR 100dB-PSRR multi-channel neural-recording amplifier system using differentially regulated rejection ratio enhancement in 0.18μm CMOS," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, San Francisco, CA, 2018, pp. 472–474. doi: 10.1109/ISSCC.2018.8310389

[11] J. Huang, T. Huang and F. Li, "Design of a low electrode offset and high CMRR instrumentation amplifier for ECG acquisition systems," *2018 14th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, Qingdao, 2018, pp. 1–3. doi: 10.1109/ICSICT.2018.8565641

[12] U. Cini, "A low-offset high CMRR current-mode instrumentation amplifier using differential difference current conveyor," *2014 21st IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Marseille, 2014, pp. 64–67. doi: 10.1109/ICECS.2014.7049922

[13] B. P. Sharma and R. Mehra, "Design of CMOS instrumentation amplifier with improved gain & CMRR for low power sensor applications," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, 2016, pp. 72–77. doi: 10.1109/NGCT.2016.7877392

[14] E. A. Azab and S. A. Mahmoud, "New CMOS realization of the differential difference operational floating amplifier with wide input voltage range," *2008 51st Midwest Symposium on Circuits and Systems, Knoxville*, TN, 2008, pp. 694–697. doi: 10.1109/MWSCAS.2008.4616894

[15] X. L. Zhang and P. K. Chan, "An untrimmed CMOS amplifier with high CMRR and low offset for sensor applications," *APCCAS 2008 - 2008 IEEE Asia Pacific Conference on Circuits and Systems*, Macao, 2008, pp. 802–805. doi: 10.1109/APCCAS.2008.4746144

[16] A. Hatim, Z. Kamal, A. Hicham and Q. Hassan, "Novel 0.064us Settling Time CMOS OP-AMP with 0.62 mW Power Consumption," *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, Fez, Morocco, 2019, pp. 1–5. doi: 10.1109/WITS.2019.8723691

[17] A. Beckers, F. Jazaeri, A. Ruffino, C. Bruschini, A. Baschirotto and C. Enz, “Cryogenic characterization of 28 nm bulk CMOS technology for quantum computing,” *2017 47th European Solid-State Device Research Conference (ESSDERC)*, Leuven, 2017, pp. 62-65. Doi: 10.1109/ESSDERC.2017.8066592

[18] R. M. Incandela, L. Song, H. Homulle, E. Charbon, A. Vladimirescu and F. Sebastiano, “Characterization and Compact Modeling of Nanometer CMOS Transistors at Deep-Cryogenic Temperatures,” *in IEEE Journal of the Electron Devices Society*, vol. 6, pp. 996-1006, 2018. Doi: 10.1109/JEDS.2018.2821763

[19] A. Suna, İ. Çevik and M. B. Yelten, “A High Speed 180 NM CMOS Cryogenic SAR ADC,” *2018 18th Mediterranean Microwave Symposium (MMS)*, Istanbul, 2018, pp. 116–119. Doi: 10.1109/MMS.2018.8611968

[20] O. V. Dvornikov, V. L. Dziatlau and N. N. Prokopenko, “Software and hardware complex for studying semiconductor devices at low, incl. cryogenic, temperatures,” *2017 2nd International Ural Conference on Measurements (UralCon)*, Chelyabinsk, 2017, pp. 253-258. Doi: 10.1109/URALCON.2017.8120719

[21] F. J. Franco, J. Lozano, J. P. Santos and J. A. Agapito, "Degradation of instrumentation amplifiers due to the nonionizing energy loss damage," *in IEEE Transactions on Nuclear Science*, vol. 50, no. 6, pp. 2433–2440, Dec. 2003. doi: 10.1109/TNS.2003.820628

[22] K. Jeong et al.," A Radiation-Hardened Instrumentation Amplifier for Sensor Readout Integrated Circuits in Nuclear Fusion Applications," *Electronics*, 2018, vol. 7, no. 12, pp. 429. doi: 10.3390/electronics7120429

[23] I. Yu. Lovshenko, V. T. Khanko, V. R. Stempitsky, " Radiation influence on electrical characteristics of complementary junction field-effect transistors exploited at low temperatures," *Materials Physics and Mechanics*, 2018, vol. 39, pp. 92–101. DOI: 10.18720/MPM.3912018_15

[24] V.A. Babenko, V.I. Gulik, L.L. Jenkovszky, V.N. Pavlovich, E.A.Pupirina," On the subcritical amplifier of neutron flux based on enriched uranium," *Nuclear Science and Safety in Europe*, Jan. 2007, pp.253-263. DOI: 10.1007/978-1-4020-4965-1_21

[25] "ISL70444SEH. Neutron Testing," InterSil, Jul. 6, 2015, pp. 1- 10. [Online]. Available: https://www.renesas.com/eu/en/doc/test-reports/neutron/isl70444seh-neutron-test-report.pdf

[26] R. A. Riedel, A. L. Wintenberg, L. G. Clonts, R. G. Cooper," High speed preamplifier circuit, detection electronics, and radiation detection systems therefrom," WO Patent appl. 2008070349 A2, Oct. 27, 2006

[27] Hibi Y. et al. “Cryogenic ultra-low power dissipation operational amplifiers with GaAs JFETs”, *Cryogenics*, 2018, vol. 73, pp. 8-13. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001122751500123X

[28] Prokopenko N.N., Dvornikov O.V., Zhuk A.A., Pakhomov I.V., "Low-temperature input stage of the operational amplifier with increased attenuation of the input common-mode signal on complementary field-effect transistors with a control p-n junction," Patent Application No. 2020104240/08, 01/31/2020

[29] Prokopenko N.N., Zhuk A.A., Pakhomov I.V., Budyakov P.S., "Low-temperature operational amplifier with increased attenuation of the input common-mode signal on complementary field-effect transistors with a control p-n junction," Patent Application No. 2020104005/08, 01/30/2020

[30] L. Kouhalvandi, S. Aygün, E. O. Güneş and M. Kırcı, "An improved 2 stage opamp with rail-to-rai! gain-boosted folded cascode input stage and monticelli rail-to-rail class AB output stage," 2017 *24th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Batumi, 2017, pp. 542–545. doi: 10.1109/ICECS.2017.8292126

[31] R. Navidi, A. Fathi, K. Mohammadi, M. Mousazadeh and A. Mousazadeh, "Improved Gain Folded Cascode OpAmp Employing a Novel Positive Feedback Structure," *2019 27th Iranian Conference on Electrical Engineering (ICEE)*, Yazd, Iran, 2019, pp. 269–273. doi: 10.1109/IranianCEE.2019.8786683

[32] A. J. Lopez-Martin, M. P. Garde, J. M. Algueta, C. A. de la Cruz Blas, R. G. Carvajal and J. Ramirez-Angulo, "Enhanced Single-Stage Folded Cascode OTA Suitable for Large Capacitive Loads," *in IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 4, pp. 441–445, April 2018. doi: 10.1109/TCSII.2017.2700060

Software Development of Electronic Digital Signature Generation at Institution Electronic Document Circulation

Nikita I. Chesnokov
Information System Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
4esnog@gmail.ru

Denis A. Korochentsev
Information System Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
mytelefon@mail.ru

Larissa V. Cherckesova
Information System Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
chia2002@inbox.ru

Olga A. Safaryan
Information System Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
safari_2006@mail.ru

Vladislav E. Chumakov
Information Systems and Radio engineering Department
Don State Technical University
Rostov–on–Don, Russia
chumakov.dssa@mail.ru

Irina A. Pilipenko
Information Systems Cybersecurity Department
Don State Technical University
Rostov–on–Don, Russia
irenphil@yandex.ru

***Abstract*— the purpose of this paper is investigation of existing approaches to formation of electronic digital signatures, as well as the possibility of software developing for electronic signature generation at electronic document circulation of institution. The article considers and analyzes the existing algorithms for generating and processing electronic signatures. Authors propose the model for documented information exchanging in institution, including cryptographic module and secure key storage, blockchain storage of electronic signatures, central web–server and web–interface. Examples of the developed software are demonstrated, and recommendations are given for its implementation, integration and using in different institutions.**

Keywords— electronic digital signature, electronic document circulation, RSA, DSA, ECDSA, GOST RF Standard 34.10–2018,secure key storage, blockchain storage of electronic signatures

I. INTRODUCTION

Information technologies and based on them information systems (IS) are widely used, at present, in the all spheres of state power authorities' activities and institutions of various forms of ownership. Information circulating in IS can be presented both in documented and undocumented forms [1]. Documented information is recorded on material carrier, with details that allow determining such information, or in the cases established by the legislation of RF its material carrier.

One of important forms of documented information is an electronic message —transmitted or received information by user of information–and–telecommunications network.

In accordance with [2], an electronic message signed with an electronic digital signature (EDS) or other analogous of handwritten signature is recognized as electronic document (ED) equivalent to the document signed with handwritten signature. Using of ED gives a number of indisputable advantages over the classic (paper) scheme of document circulation, namely:

- Significant reduction in time spent on document transfer;
- Reduce the number of errors in documents, as well as simplify the correction of detected errors;
- Saving on consumables and postal services;
- Eliminating the need to maintain paper archives that take up a lot of space, as well as creating the potential for documents loss.

One of the first mentions of electronic signature appeared in 1976 in the work "New Directions in Cryptography", which described, based on classical cryptographic systems, the possibility of creating fully electronic digital mechanism similar to handwritten signature [3].

A year later, Ronald Rivest, Adi Shamir and Leonard Adleman developed the well–known and still frequently used RSA cryptographic algorithm. Their article "A Method for Obtaining Digital Signatures and Public–Key Cryptosystems" describes this algorithm, as well as the method of its application for generating of digital signatures [4].

These investigations served as impetus for further theoretical and practical development of cryptography in the field of ES. After the RSA algorithm, other digital signature schemes were developed, such as the DSA, ECDSA, Rabin and Merkle algorithms and others [5], [6], and [7].

The history of separate development of ES in Russia began in 1994, with the development of the first Russian EDS standard GOST R 34.10–94 by the main department of communication security "Federal Agency for Government Communications and Information" (FAGCI) [8]. In the future, the regulations in the

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

field of application of ES were finalized [9, 10], and, at present, GOST Standard 34.10–2018 is relevant and actual [11].

Currently, there is significant growth of the ES market in Russia, which is directly due to increasing in the number of issued ES certificates. Over the period from 2006 to 2015, this market grew from 1.5 billion rubles to 14.8 billion rubles that is almost in 9.9 times [12]. Along with the growing interest in the legitimate using of electronic signatures, the incentive interest of hackers is also growing. In 2019, it became known about the increasing number of cases of fraudulent transactions with EDS using. All these enumerated facts indicate that the electronic digital signature is promising and relevant, and since it is currently undergoing active development and implementation, the number of tasks of various kinds associated with it will only grow in the near future.

Scientific novelty of the proposed work – the application of blockchain technology for the remote cloud formation and verification of electronic digital signature. Blockchain storage is necessary for the reliable storage of electronic keys that can be obtained includingPKI(Public KeyInfrastructure) technology use. Purpose of this work is to create software tool for remote (cloud) generation and verification of EDS using blockchain technology. The goal achieving involves the next scientific tasks:

– Study of juridical base for electronic signatures application;
– Review and analysis of existing similar software products;
– Development of architecture and algorithms of software tool;
– Practical implementation and testing of developed software.

II. High Level Software Architecture

The components of created complex are implemented in Python programming language, version 3.5; development environment is Microsoft Visual Studio Code. The graphical interface of the software tool is implemented as web page, programming language is JavaScript; hypertext markup language is HTML; and style sheet is CSS. This approach will ensure maximum coverage of supported devices, since any modern computer and smartphone allow viewing web pages.

As type of electronic digital signature is used the simple one. The advantages of choosing blockchain technology for remote cloud generation and verification of electronic digital signatures in comparison with the traditional chain of trust of certificates with a digital signature are the presence of a blockchain storage, which is used for reliable storage of electronic keys, which, in turn, can be obtained, including using Public Key Infrastructure (PKI) technology.

An open storage of generated signatures is used as component increasing the level of trust to the cloud scheme of signature generation. Blockchain technology is implemented to provide protection against integrity violations, as well as to ensure that the reliable time of its filling, or, in other words, the time of signature formation, can be applied. The proposed architecture of the developed software complex is presented graphically in Fig. 1.

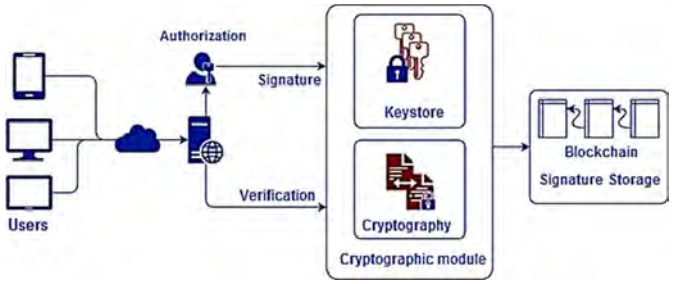


Fig. 1. Architecture of developed software package

It includes the following parts:

– Secure cryptographic module that performs cryptographic operations with documents, signatures, and signature keys;

– Secure signature key storage that is an integral part of the cryptographic module;
– Blockchain storage – the repository of generated signatures;
– Central web–server;
– User's web–interface.

III. Cryptographic Module and Secure KeyStorage

Private keys for signatures are stored securely on the service side, "in the cloud". Key security is supported by encryption, which encrypts both the entire storage and each key separately. Only the owner, who has passed the identification, authentication, and authorization procedures, has access to particular key.

Authorized users do not have access to private keys, i.e. even key owners initiate the creation of their keys upon registration in the system, but do not receive the keys themselves. They can control their use by authorization into the system only.

The document signing procedure is performed "in the cloud" also, in the cryptographic module (Fig. 2). To sign the document, the user must pass the authorization procedure, after which the software tool generates electronic digital signature (EDS) that is used to sign the document.

The generated signature is not only sent to the user, but also added to the blockchain signature storage. In addition to the bit sequence of the signature itself, certificate of its public key is added to the blockchain storage, which contains data about the signature owner, as well as its validity period.

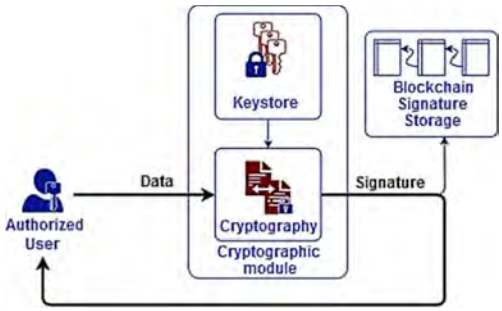


Fig. 2. Scheme of the electronic digital signature forming and signing an electronic document

Both in the electronic signatures themselves and in ES– blocks store the information about the time of creation, which allows increasing the confidence level to the developed system. For verification, in addition to the usual procedure for hash functions comparing, the signature uploaded by the user is determined in the blockchain storage, and only if it is found, the signature is considered as correct and legitimately formed. If the signature turns out to be correct according to the usual verification procedure, but is not present in signature store, this fact serves as automatic signal of possible defamation of system.

IV. Blockchain Signature Storage

Blockchain, which means "chain of blocks", is structure that has all the properties of connected list, as well as a number of additional properties. Namely, each block is associated with all previous ones by certain regularity, and it is forming by solving of certain problem, aimed at compliance with low, and requiring certain amount of computing resources to solve it [13, 14].

If this regularity is violating in particular block, it and all subsequent blocks are considered as invalid, since each bit of block information is involved in forming the hash functions of all subsequent blocks. This means that changing of any bit in the N–2 block of chain from N blocks will change the hash functions of N–2, N–1, and N blocks.

The considered architecture of the blockchain storage ensures its protection from unauthorized changes and the integrity of the electronic digital signature.

The structure of the blockchain signature storage used in the developed software system is shown graphically in the figure 3.

As Fig. 3 demonstrates, each block can contain unlimited number of signatures, but not zero. It also must contain its own information, such as its own hash function, the hash function of the previous block, the impurity (salt) for the forming of its own hash function, the date of formation, and the identifier.

V. Central Web Server and Web Interface

All components of the system are developed independently; interact with each other through http–requests using the REST programming interface. This provides the ability to scale the system that is necessary, first of all, for constructing of distributed blockchain storage of signatures.

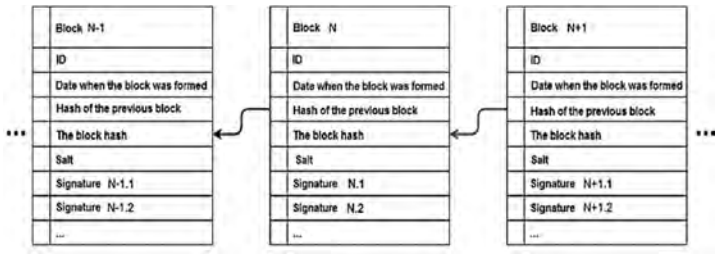


Fig. 3. Structure of the blockchain signature repository

Central web server provides the interaction of all elements of the system. It processes the user's http–requests for registration, signing and verification of signatures, and after processing, redirects the received data and requests to the cryptographic module or the blockchain storage of signatures.

From the user's point of view, all interaction with the system comes down to filling out some form on the web page. To verify the signature, it is enough to download the signed file and signature; and for signing, in addition to downloading the document, the user also need to go through the authorization procedure that is, to enter the login and password to "indicate" the service the electronic key of the signature and confirm it ownership.

VI. Software Demonstration

Demonstration of the work of electronic signature generation software held for organizing electronic document circulation in the institution.

The software starts using the console command, which in general can be represented as follows:

python main .py -a {rsa, dsa, gost} -k {key}.

The first argument (-a or --algo) takes one of three possible values that correspond to the supported electronic signature algorithms (RSA, DSA, GOST 34.10–2018), and the second argument (-k or --key) is the encryption key that encrypts the signature key store.

In Fig. 4, the start of software tool is presented graphically, in the mode of working with RSA algorithm and the encryption key of the block storage of signatures:

d8f0e535fccad9fd4ee11cfd3cf1138cd9d53e13bebddcd2daf993c47c014b8e.

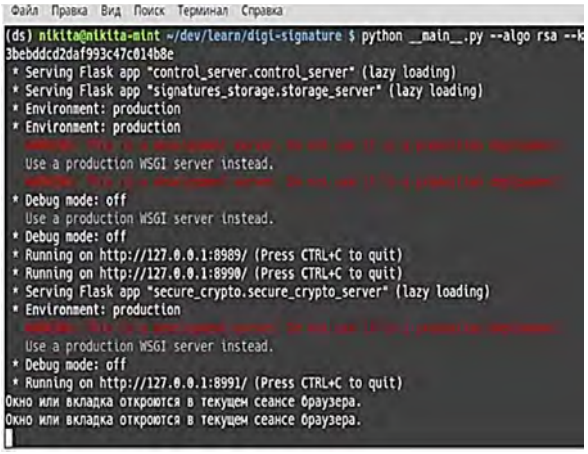


Fig. 4. Start of the electronic signature generation software

This command starts a program that processes the input arguments and then generates three subprocesses: the central web server process, the cryptographic module process, and the blockchain storage of signatures process. It is possible to implement software tool in which each of the specified subprocesses can be started separately, for example, on the physically different computers.

The software will open automatically the graphical user interface in the new browser window. If no users are registered in the system, the graphical administrator interface is called at address / admin (this situation is shown in the Fig. 5).

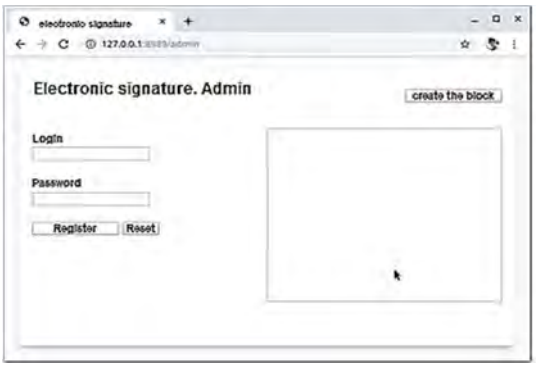


Fig. 5. Graphical interface of administration panel, creating new user

When user registering, the administrator enters the user's username and password, then clicks the button "Register". After successful registration, a corresponding message will appear in the text field on the right side of the interface, and the link will appear under the login and password entry form to download the certificate of the signature verification key of the newly created electronic signature key (Fig. 6).

Fig. 6. Registering the new user of the software tool

Functionally, the administration panel provides for the forced formation of new block in the blockchain signature storage (the "Create the block" button).

When viewing the generated electronic signature key in the key store, the data in the storage is not readable, which allows ensuring their confidentiality even in the case when the key is compromised (Fig. 7).

Fig. 7. Encrypted signature key storage

Web interface of user is represented graphically in Fig. 8. It includes text fields for entering of username and password, the switch of operating modes, the field for loading of signed data, and the ED signature function (the "Sign" button).

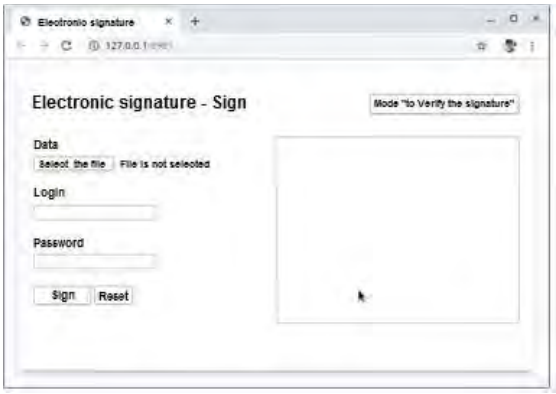


Fig. 8. Web user interface

When the form fills out correctly (specifying the actual login and password of the user), after clicking the "Sign" button, the message about the successful signing of electronic document and link to download the generated signature are displayed, which is graphically illustrated in the Fig. 9.

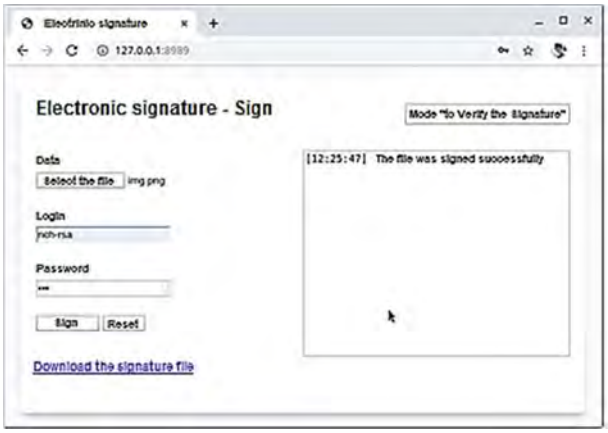


Fig. 9. Successful file signing img.png

Generated electronic signature is not only sent to user but is saved automatically in the blockchain storage also (Fig. 10).

Fig. 10. Contents of the blockchain signature repository

To verify the legitimacy of electronic signature, it is necessary to verify it. For this purpose, user's web interface can be switched to the mode "Verify the signature". After downloading the previously signed document and the downloaded electronic signature, the signature check must be running. If it is successful, in the right part of interface will display the conclusion about the validity or invalidity of the signature, as illustrated in Fig. 11.

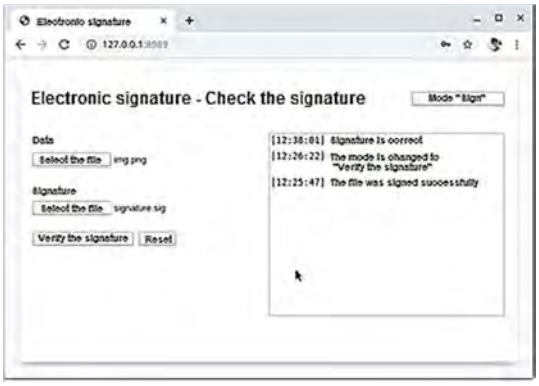


Fig. 11. User's web interface in EDS verification mode

VII. Conclusion

The principal difference is the application of additional security module (reliable blockchain key storage), which increases the crypto stability of the system as a whole.

The software for electronic signature creating implements a cloud–based signature storage scheme, and includes an additional element–blockchain storage of electronic signatures generated by system. It designed to increase the confidence level both in the developed software as a whole, and in each electronic signature generated using it separately.

It is quite difficult to compare the developed system with its analogues, since the authors are the first to use blockchain technology for remote cloud generation and verification of electronic signatures, and have not found the similar approach in other EDS–systems. In addition, electronic digital signature systems are mostly commercial, and it is problematic to get access to them for their functional research.

The considered software tool includes: the user interface, the central web server, the cryptographic module that includes modules for cryptographic operations and secure key store for electronic signatures, as well as separate module for the blockchain storage of generated electronic signatures.

All of these elements are designed with scalability and extensibility in mind. They can work independently of each other. Communication between them occurs via http–requests. As part of demonstration of the full cycle of working with the software, main functions implemented by it were shown. The developed software can be used at the organization of electronic document circulation in any form of ownership that allows the use of simple electronic signature.

References

[1] "On information, information technologies and information protection" – Federal law of Russian Federation No. 149–FZ of July 27, 2006 (In Russian).

[2] "On electronic signature" – Federal law of Russian Federation No. 63-FZ of April 06, 2011 (In Russian).

[3] W. Diffie, M. Hellman, New "Directions in Cryptography". IEEE Trans. Inf. Theory. Kschischang. IEEE, 1976. Vol. 22, Iss. 6. Pp. 644 – 654.Doi:10.1109/TIT.1976.1055638.

[4] R. Rivest, A. Shamir, L. Adleman, "Method for obtaining digital signatures and public key cryptosystems". Commun. ACM. New York City: ACM, 1978. Vol. 21, Iss. 2. Pp. 120–126. Doi:10.1145/359340.359342.

[5] R. Merkle, "Secrecy, authentication, and public key systems". Technical Report No. 1979. Citeseer, 1979. Doi: 10.11.637.3952.

[6] FIPS PUB 186–4. Federal Information Processing Standards Publication. Digital signature standard (DSS). Iss. 2013–07. Gaithersburg,MD:NIST, 2013.Doi:10.6028/NIST.FIPS.186–4.

[7] ANSI X 9.62. "Public Key Cryptography for Financial Services Industry: TheEllipticCurve DigitalSignature Algorithm(ECDSA)". Iss. 2005–11–16. ANSI, 2005.

[8] GOST R 34.10–94. Informacionnaja tehnologija. Kriptograficheskaja zashchita informacii. Procedury vyrabotki i proverki elektronnoj cifrovoj podpisi na baze asimmetrichnogo kriptograficheskogo algoritma. Vved. 1995–01–01. M.: Izd–vo standartov, 1995. 9 p.

[9] GOSTR34.10–2001.Informacionnajatehnologija. Kriptograficheskaja zashchita informacii. Processy formirovanija i proverki jelektronnoj cifrovoj podpisi.Vved. 2002–001.M.: Izd-vo standartov, 2001.12 p.

[10] GOSTR34.10–2012.Informacionnajatehnologija. Kriptograficheskaja zashchita informacii. Processy formirovanija i proverki jelektronnoj cifrovoj podpisi. Vved. 2013–01–01. M.: Standartinform, 2013.16 p.

[11] GOST 34.10–2018. Informacionnaja tehnologija. Kriptograficheskaja zashchita informacii. Processy formirovanija i proverki jelektronnoj cifrovoj podpisi. Vved. 2019–06–01. M.: Standartinform, 2018. 15 p.

[12] K. Bokieva, "Eelektronnajapodpis' v Rossii: sostojaniei perspektivy". Molodoj uchenyj. 2018. № 52. Pp. 185–186. URL: https://moluch.ru/archive/238/55256 (data obrashhenija:21.11.2019).

[13] D. Drescher, "Fundamentals of blockchain: an introductory course for beginners in 25 small chapters". DMK Press, 2017. 320 p.

[14] R. Wattenhofer, "The Science of the Blockchain". Inverted Forest Publishing, 2016, 123 p.

Evaluating Length of a Shortest Adaptive Homing Sequence for Weakly Initialized FSMs

Nina Yevtushenko
Software engineering deparment
Ivannikov Institute for System Programming of RAS
Moscow, Russia
nyevtush@gmail.com

Evgenii Vinarskii
Computer science department
Lomonosov Moscow State University
Moscow, Russia
vinevg2015@gmail.com

***Abstract*—— There are many research papers devoted to the state identification problem of finite state machines (FSMs) which are widely used for analysis of discrete event systems. A homing sequence (HS) always exists for a deterministic complete reduced FSM but a number of non-deterministic FSMs not necessarily have a preset HS; nevertheless, those FSMs can still have an adaptive HS. Adaptive sequences exist more often and can be shorter than the preset. If each state can be an initial state of a complete FSM, i.e., an FSM is non-initialized, then a shortest adaptive HS (if it exists) has length that is polynomial with respect to the number of FSM states. In this paper, we show that it is not the case for weakly initialized FSMs; for such FSMs, the length of a shortest adaptive HS can be exponential in the number of FSM states / transitions. However, the experimental evaluation shows that the non-polynomial upper bound for a shortest adaptive HS was never reached for randomly generated weakly initialized two-input FSMs.**

Keywords—Non-deterministic Finite State Machine (FSM), weakly initialized FSM, adaptive homing sequence

I. INTRODUCTION

Distinguishing, Homing, Synchronizing (DS, HS, SS) sequences of Finite State Machines (FSM) are widely used for the optimization of test suites for telecommunication protocols and other discrete event systems [1-7]. These sequences are useful for minimizing a test suite with guaranteed fault coverage as well as for reducing monitoring efforts by minimizing sets of invariants to be checked [8, 9]. Preset state identification sequences are derived in advance, while for adaptive state identification sequences the next input essentially depends on the outputs produced to the previous inputs. An adaptive input sequence is usually represented as a tree or an acyclic initialized FSM (a test case [10-12]). For deriving preset DS and HS, an appropriate distinguishing or homing tree is constructed while when deriving adaptive sequences, a distinguishing or homing FSM [13-15] is utilized. A different QBF based approach for deriving a preset homing sequence is proposed in [16]. According to the performed experiments with randomly generated non-deterministic FSMs [17] adaptive distinguishing and homing sequences exist more often and are usually shorter than the preset.

A homing sequence allows drawing a conclusion about a current state of an FSM under experiment independently of the initial state of the machine. For a complete deterministic FSM that is reduced, strongly connected and non-initialized, i.e. a machine where each state can be an initial state, there always exists a preset homing sequence and its length is polynomial in the number of FSM states but it is not the case for a non-deterministic FSM, i.e., a preset HS may not exist for such FSMs. Moreover, given a complete non-deterministic FSM, a shortest preset homing sequence can have exponential length in the number of FSM states [18]. There exist necessary and sufficient conditions for checking the existence and deriving preset and adaptive homing sequences for complete and partial non-initialized non-deterministic FSMs [19, 20, 21]. In particular, an adaptive homing sequence exists for a complete non-initialized FSM if and only if such a homing sequence exists for each pair of different FSM states. This fact implies that the existence check as well as the derivation of an adaptive homing sequence is in **P** [13, 19]. For weakly initialized FSMs the above condition becomes only sufficient and a number of heuristics is proposed for deriving a homing sequence of reasonable length [see, for example, 22]. In one of the approaches [12] homing subsets of states are enumerated starting from state pairs until the set of initial states is obtained or no homing subsets can be further derived; this approach becomes rather complicated for machines with many states. In [17], homing FSMs are utilized for deriving adaptive homing sequences for weakly initialized FSMs. However, there is no evaluation of length of a shortest adaptive homing sequence when such a sequence exists.

In this paper, we prove that length L of a shortest adaptive homing sequence for a weakly initialized observable non-deterministic FSM can be exponential in the number of FSM states / transitions, i.e., with respect to the FSM size, and thus, the check of the existence of an adaptive homing sequence is at least NP-hard. The latter almost immediately implies the same result for non-observable non-initialized complete FSMs. In fact, we use the same strategy as in [23] for constructing a class of observable complete weakly initialized FSMs for which L is exponential with respect to the FSM size.

The rest of the paper has the following structure. The preliminaries are in Section II. A class of observable complete weakly initialized FSMs for which a shortest adaptive homing sequence has exponential length in the number of FSM states / transitions is described in Section III and Section IV concludes the paper.

II. PRELIMINARIES

In this section, we briefly remind the main definitions and notations mostly taken from the papers [11, 17].

A. Finite State Machines

Finite state machines (FSM), or simply *machines*, are widely used for analysis and synthesis of discrete event (digital) systems. Formally, an FSM $\mathsf{S} = \langle S, I, O, h_{\mathsf{S}}, S_{in}\rangle$ is a 5-tuple with a finite nonempty set S of states and a set $S_{in} \subseteq S$ of initial states, finite input and output alphabets I and O, and a *transition relation* $h_{\mathsf{S}} \subseteq S \times I \times O \times S$. If $S_{in} = S$ then FSM S is *non-initialized*; if $|S_{in}| = 1$ then FSM S is an *initialized* FSM; if $1< |S_{in}| < |S|$ then FSM S is *weakly initialized*. If for some pair $(s, i) \in S \times I$, there exist at least two different pairs (o', s'), $(o'', s'') \in O \times S$ such that (s, i, o', s'), $(s, i, o'', s'') \in h_{\mathsf{S}}$ then FSM S is *non-deterministic*; otherwise, the FSM is *deterministic*. A non-deterministic FSM S is *observable* if for every two transitions (s, i, o, s_1), $(s, i, o, s_2) \in h_{\mathsf{S}}$ it holds that $s_1 = s_2$; otherwise, the machine is *non-observable*. FSM S is *complete* if for every pair $(s, i) \in S \times I$, there exists a transition $(s, i, o, s') \in h_{\mathsf{S}}$. A flow table of a complete nondeterministic observable FSM is shown in Figure 1 where lines correspond to inputs while columns correspond to states of the machine. A corresponding unit for state s and input i has a pair s'/o if $(s, i, o, s') \in h_{\mathsf{S}}$. The FSM has seven states, $S = \{1, 2, \ldots, 7\}$, two inputs, $I = \{x_1, x_2\}$, and four outputs, $O = \{0, 1, 2, 3\}$. By definition, the FSM is complete, observable and non-deterministic. For example, given a state 1 and input x_2, there are two transitions in the FSM: $(1, x_2, 1, 1)$ and $(1, x_2, 2, 6)$. Given a complete FSM S, $io \in IO$ and $s \in S$, the io-successor of state s has each state s' of FSM S such that $(s, i, o, s') \in h_S$. If the io-successor of state s is empty then we sometimes say that the io-successor of state s *does not exist*.

An FSM is a sequential model and correspondingly, we extend the behavior relation h_{S} to input and output sequences. A sequence $i_1o_1 \ldots i_lo_l$ (or $i_1/o_1 \ldots i_l/o_l$) of input/output pairs labeling successive transitions starting from state s is a *trace* of FSM S at state s. Given a state s and an input sequence $i_1 \ldots i_l$, an output sequence $o_1 \ldots o_l \in out(s, i_1 \ldots i_l)$ if and only if $i_1o_1 \ldots i_lo_l$ is a trace at state s. When FSM S is observable, for each trace γ of the FSM, it holds that the γ-successor of state s either does not exist or is a singleton $\{s'\}$; in the latter case, s' is a state where γ takes the FSM from state s. If FSM S is non-observable then several states can be reached from s via γ and thus, the γ-successor of state s not necessarily is a singleton. If S' is a non-empty subset of states of the FSM, then the union of γ-successors over all states of the set S' is the γ-successor of S'. The FSM is *strongly connected* if for each two different states s and s', state s' is reachable from s. In this paper, we analyze the length of homing sequences for complete and observable weakly initialized FSMs.

Test case definition. An *adaptive* input sequence when the next input depends on an output to the previous input can be represented by a proper tree or a special FSM called a *test case* [10-12]. A *test case* $\mathsf{TC}(I, O)$ over input and output alphabets I and O is an initialized observable initially connected acyclic FSM. Moreover, at each state of the test case either only one input is defined with all possible outputs or the state has no outgoing transitions. Consider an FSM with the flow table in Figure 1 and a test case in Figure 2 in order to illustrate how a test case represents an adaptive experiment with a complete FSM over alphabets I and O.

Given a test case $\mathsf{TC}(I, O)$, $I = \{x_1, x_2\}$, $O = \{0, 1, 2, 3\}$, the test case describes an adaptive experiment with a complete FSM S_1 over alphabets I and O when the set of initial states has two states 1 and 3. At the first step, the input x_1 defined at the initial state $\{1, 3\}$ of $\mathsf{TC}(I, O)$ is applied to the FSM S_1 in Figure 1. If S_1 produces output 3 as the response to the input x_1, then it is known that the FSM reached state 5 and the experiment is terminated. If S_1 produces the output 0 as the response to the input x_1, then the next state of $\mathsf{TC}(I, O)$ is the x_1o-successor $\{2, 4\}$ of the state $\{1, 3\}$. The input x_1 defined at state $\{2, 4\}$ is applied at the next step, etc. Once the test case reaches a deadlock state, the procedure is over. For simplicity, in Figure 2, we do not show transitions to deadlock states which cannot appear for the given FSM. Length of a longest trace from the initial state to a deadlock state of $\mathsf{TC}(I, O)$ is the *height* of the test case; in other words, the test case height specifies the maximum length of an input sequence that can be applied to an FSM under experiment. In our example, the test case height is seven.

Given a complete weakly initialized FSM $\mathsf{S} = \langle S, I, O, h_{\mathsf{S}}, S_{in}\rangle$, a test case $\mathsf{TC}(I, O)$ is a *homing* test case (HTC) for S if for every trace γ from the initial state to a state without outgoing transitions, the γ-successor of the set S_{in} in S is a singleton or the empty set. If there exists an HTC for S then the FSM is *homing*. By direct inspection, one can assure that the test case in Figure 2 is an HTC for FSM S_1 in Figure 1. A homing test case is a *synchronizing* test case (STC) for S if there exists a state s such that for every trace γ from the initial state to a state without outgoing transitions, the γ-successor of the set S_{in} in S is a singleton $\{s\}$ or the empty set. A test case in Figure 2 is not an STC for FSM S_1.

St/ inputs	1	2	3	4	5	6	7
x_1	2/0 5/3	1/0	4/0 2/3	5/0	3/0	6/0 2/3	6/0 2/3
x_2	1/1 6/2	5/1 7/2	3/1 6/2	4/1 6/2	5/1 7/2	6/2	7/2

Fig. 1. The flow table of FSM S_1

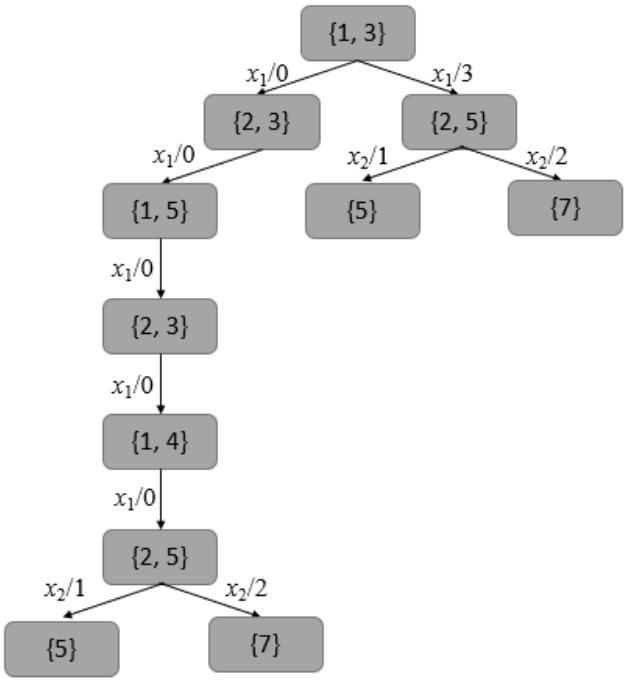


Fig 2. HTC T for the FSM S_1

Given a shortest preset homing sequence for a non-deterministic FSM, it can happen that its length is exponential with respect to the number of FSM states / transitions [18]. In this paper, we show that this can happen for an adaptive homing sequence when an FSM is weakly initialized, i.e., there exists a class of complete observable FSMs for which the height of a shortest HTC is exponential with respect to the size of a nondeterministic FSM under experiment.

It is known how to derive a shortest homing test case for non-initialized and weakly initialized FSMs. A non-initialized FSM is homing if and only if each pair of different states has an HTC. Correspondingly, the existence check and derivation of a homing test case for a complete observable non-initialized FSM have the polynomial complexity [12, 19]. When an FSM is weakly initialized, the above condition becomes only sufficient. Another necessary and sufficient conditions are based on deriving an appropriate homing FSM [17] but for this approach, the complexity of the existence check and derivation of a homing test case is not evaluated as well as the height of a shortest homing test case.

In this paper, we show that height of a shortest homing test case that is the length of a shortest adaptive homing sequence for a weakly initialized complete observable FSM can be exponential with respect to the number of states / transitions of an FSM under experiment, i.e., with respect to the FSM size. Correspondingly, the problem of the existence check of a homing test case for a weakly initialized complete FSM is at least NP-hard and this is a hint that the problem of the existence check of an adaptive synchronizing sequence for a weakly initialized FSM is also at least NP-hard.

III. Evaluating Length of a Shortest Adaptive Homing Sequence

In this paper, we are concerned about homing test cases (HTC) for complete observable weakly initialized non-deterministic FSMs in order to investigate whether the height of a shortest HTC can be exponential in the number of FSM states / transitions and show that in fact, this is the case. The latter allows to conclude that the problem of checking the existence of an HTC for a complete observable weakly initialized FSM is at least NP-hard.

In order to derive an appropriate class of FSMs with the above property, we use the same approach as in [23] when evaluating length of a shortest adaptive distinguishing sequence for complete observable non-deterministic FSMs. We show that for any integer n, one can construct a two-input FSM S such that the size of FSM S is polynomial in n, but the minimal height of a homing test case for an appropriate subset of initial states of S is exponential in n.

Let $n \geq 2$ be an integer such that $p_1, p_2, \ldots, p_n$ are the first n different primes considered in increasing order. Furthermore, let $\Sigma_n = p_1 + p_2 + \ldots + p_n$ be the sum of the first n primes, and let $\Pi_n = p_1 \times p_2 \times \ldots \times p_n$ be the product of the first n primes. We show that there exists an FSM S_n with $\Sigma_n + 2$ states such that the minimal length of an adaptive homing test case for an appropriate subset of initial states equals $\Pi_n + 2$. Note that Σ_n is polynomial in n and Π_n is exponential in n. The state set of the FSM S_n is $S = \{1, 2, \ldots, \Sigma_n, \Sigma_n + 1, \Sigma_n + 2\}$ while the set of initial states is $\{1, \Sigma_1 + 1, \ldots \Sigma_{n-1} + 1\}$. We consider the set of states partitioned into n subsets $S_1, S_2, \ldots, S_n$, where $S_j = \{\Sigma_j - p_j + 1, \Sigma_j - p_j + 2, \ldots, \Sigma_j\}$, for $1 \leq j \leq n$. An FSM has two inputs x_1 and x_2 and the set of outputs is $\{0, 1, 2, 3\}$. The transitions under x_1 constitute a cycle of length p_j for the states in S_j, for $1 \leq j \leq n$, with the same output 0. Formally, for a state $k \in S_j$, for $1 \leq j \leq n$, we have the transition $(k, x_1, 0, K)$ where $K = k + 1$ if $k < \Sigma_j$ and $K = \Sigma_j - p_j + 1$ if $k = \Sigma_j$. For a state $k \in S_j$, for $1 \leq j \leq n$, we have transitions $(k, x_2, 1, k)$ and $(k, x_2, 2, \Sigma_n + 1)$ if $k < \Sigma_j$, the transitions $(\Sigma_j, x_2, 1, \Sigma_{j+1})$ and $(k, x_2, 2, \Sigma_n + 2)$ if $j < n$, the transitions $(\Sigma_n, x_2, 1, \Sigma_n)$ and $(\Sigma_n, x_2, 2, \Sigma_n + 2)$ and the transitions $(\Sigma_n + 1, x_2, 2, \Sigma_n + 1)$ and $(\Sigma_n + 2, x_2, 2, \Sigma_n + 2)$. For output 3, there also are transitions $(\Sigma_j + 1, x_1, 3, \Sigma_{j+1})$ for $1 \leq j \leq n - 1$, transition $(\Sigma_n + 1, x_2, 3, \Sigma_1)$ and transitions $(\Sigma_n + 1, x_1, 3, \Sigma_1)$ and $(\Sigma_n + 2, x_1, 3, \Sigma_1)$. We further show that transitions under x_2 home only states of the subset $b = \{\Sigma_1, \Sigma_2, \ldots, \Sigma_n\}$. The flow table of such an FSM for $n = 2$ is shown in Figure 1 while in Figure 2, such flow table is shown for $n = 3$. The number of states of the FSM for $n = 2$ is $\Sigma_2 + 2 = 2 + 3 + 2 = 7$ and correspondingly, for $n = 3$ the number of states is $\Sigma_3 + 2 = 2 + 3 + 5 + 2 = 12$. It can be noted that FSMs with the flow tables in Figures 1 and 3 are strongly connected.

	S_1		S_2			S_3					11	12
	1	2	3	4	5	6	7	8	9	10	11	12
x_1	2/0 5/3	1/0	4/0 10/ 3	5/0	3/0	7/0 2/3	8/0	9/0	10/ 0	6/0	11/ 0 2/3	12/ 0 2/3
x_2	1/1 11/ 2	5/1 12/ 2	3/1 11/ 2	4/1 11/ 2	10/ 1 12/ 2	6/1 11/ 2	7/1 11/ 2	8/1 11/ 2	9/1 11/ 2	10/ 1 12/ 2	11/ 2	12/ 2

Fig. 3. The flow table of FSM S_n for $n = 3$

We first establish a simple statement about state subsets of a complete observable FSM which have no HTC.

Proposition 1. 1. Given an FSM and a subset of its states, the subset has no HTC if it is not a singleton and for every input i there exists an output o such that the io-successor of the subset equals this subset. 2. Given an FSM and a subset of its states, the subset has no HTC if for every input i there exists an output o such that the io-successor of the subset has no HTC.

Indeed, if 1) holds then for any test case there exists a trace from the initial state to a deadlock state that it is not a singleton. If 2) holds then the same it valid, since otherwise 2) does not hold.

As a corollary to the above proposition, some statements about state subsets of the FSM S_n can be established.

Proposition 2. 1. Any subset of states of the FSM S_n that has states $\Sigma_n + 1$ and $\Sigma_n + 2$ has no HTC. 2. Any subset that has two different states of some Σ_j, $j = 1, \ldots, n$, has no HTC.

The statement 1) is a direct corollary to Proposition 1, since $x_1$0-successor of $\{\Sigma_n + 1, \Sigma_n + 2\}$ as well as $x_2$2-successor of $\{\Sigma_n + 1, \Sigma_n + 2\}$ equals $\{\Sigma_n + 1, \Sigma_n + 2\}$. For proving 2), consider two different states s and s' of some S_j. When using x_1, this pair of states can only reach another pair of different states of the same subset S_j. If input x_2 is used and one of states is Σ_j, the $x_2$2-successor of the pair is $\{\Sigma_n + 1, \Sigma_n + 2\}$ that has no HTC. If input x_2 is used and none of states is Σ_j, the $x_2$1-successor of the pair is $\{s, s'\}$ and thus, this pair also has no HTC.

Proposition 3. Given a subset $\{k_1, \ldots, k_n\}$, $k_1 \in S_1, \ldots, k_n \in S_n$, a shortest (adaptive) homing sequence is $(x_1)^l(x_2)^{n-1}$ where l is a minimal number such that the $x_1$0-successor of $\{k_1, \ldots, k_n\}$ equals $\{\Sigma_1, \ldots, \Sigma_n\}$.

Proof. We prove the statement into three steps P1, P2 and P3 below.

P1. Consider two states k_1 and k_2 of different S_j, $j = 1, \ldots, n$, where $k_1 \neq \Sigma_1$ or $k_2 \neq \Sigma_2$. Without loss of generality, let $k_1 \in S_1$ and $k_2 \in S_2$. If $k_1 \neq \Sigma_1$ and $k_2 \neq \Sigma_2$, then the $x_2$1-successor of $\{k_1, k_2\}$ equals $\{k_1, k_2\}$. If there exists an HTC for this pair where the first input is x_2 then the test case after $x_2$1 is an HTC for $\{k_1, k_2\}$, i.e., a shortest adaptive homing sequence cannot be headed with x_2. Consider a pair $\{k_1, \Sigma_2\}$ where $k_1 \neq \Sigma_1$. By definition, the $x_2$2-successor of $\{k_1, \Sigma_2\}$ equals $\{\Sigma_n + 1, \Sigma_n + 2\}$ that cannot be homed (Proposition 2) and thus, a pair $\{k_1, \Sigma_2\}$ cannot be homed by an adaptive sequence headed with x_2. The same situation occurs for the pair $\{\Sigma_1, k_2\}$ if $k_2 \neq \Sigma_2$. Therefore, given two states k_1 and k_2 of different S_j, $j = 1, \ldots, n$, where $k_1 \neq \Sigma_1$ or $k_2 \neq \Sigma_2$, a shortest adaptive homing sequence for the subset $\{k_1, k_2\}$ cannot be headed with the input x_2.

P2. At the next step, we note that by direct inspection, one can assure that due to construction, a subset $\{\Sigma_1, \ldots, \Sigma_n\}$ can be homed by $(x_2)^l$ if and only if $l \geq n - 1$. Indeed, the $x_2$1-successor of $\{\Sigma_{n-1}, \Sigma_n\}$ equals $\{\Sigma_n\}$, the $x_2$1-successor of $\{\Sigma_{n-2}, \Sigma_{n-1}, \Sigma_n\}$ equals $\{\Sigma_{n-1}, \Sigma_n\}$,…, the $x_2$1-successor of $\{\Sigma_1, \ldots, \Sigma_n\}$ equals $\{\Sigma_2, \ldots, \Sigma_n\}$ while the $x_2$2-successor of $\{\Sigma_1, \ldots, \Sigma_n\}$ equals $\{\Sigma_n + 2\}$. We also notice that the $x_1$3-successor of the set of initial states is $\{\Sigma_1, \ldots, \Sigma_n\}$ and thus, output 3 does not affect the height of a shortest homing test case.

P3. Consider now a subset $\{k_1, \ldots, k_n\}$, $k_1 \in S_1, \ldots, k_n \in S_n$. Due to P2, a shortest homing sequence for the subset $\{k_1, \ldots, k_n\}$ is $(x_1)^l x_2 \alpha$ where $l > 0$. If the $(x_1 0)^l$-successor $\{k'_1, \ldots, k'_n\}$ of $\{k_1, \ldots, k_n\}$ is not equal to $\{\Sigma_1, \ldots, \Sigma_n\}$ then again $x_2\alpha$ is not a shortest homing sequence for $\{k'_1, \ldots, k'_n\}$. Therefore, for a shortest adaptive homing sequence, the $(x_1 0)^l$-successor of $\{k_1, \ldots, k_n\}$ has to be $\{\Sigma_1, \ldots, \Sigma_n\}$ and $(x_1)^l(x_2)^{n-1}$ where l is a minimal number such that the $x_1$0-successor of $\{k_1, \ldots, k_n\}$ equals $\{\Sigma_1, \ldots, \Sigma_n\}$ is a shortest homing sequence for $\{k_1, \ldots, k_n\}$.

Consider a weakly initialized FSM S_n where the set of initial states is a subset $\{1, \Sigma_1 + 1, \ldots \Sigma_{n-1} + 1\}$. As a corollary to Proposition 3, the following statements hold.

Proposition 4. The minimum height of an HTC for S_n is $\Pi_n - 1 + (n - 1)$.

Theorem 1. Given $n > 1$, let $p_1, p_2, \ldots, p_n$ be the first n different primes considered in increasing order, $\Sigma_n = p_1 + p_2 + \ldots + p_n$ and $\Pi_n = p_1 \times p_2 \times \ldots \times p_n$. There exists an FSM S_n with Σ_n states and two inputs such that the minimum height of an HTC is $\Pi_n + n - 2$.

Theorem 2. Given $n > 1$, there exists a complete observable weakly initialized FSM S_n such that the minimum height of an HTC is not polynomial in the number of states of the FSM.

Note that if $p_1(x)$ is polynomial in x and $p_2(x)$ is polynomial in x then $p(x) = p_1(p_2(x))$. Indeed, if we assume that Π_n is polynomial in Σ_n, then Π_n is also polynomial in n, since Σ_n is polynomial in n [24]. Since Π_n is known to be exponential in n, the obtained contradiction proves the statement.

We also notice that if we use algorithms from [17] for deriving a homing test case, the number of states of a homing FSM S_{homing} becomes exponential. On the other hand, when deriving an HTC with minimum height, until L reaches the value $\Pi_n + n - 1$, there always exists a complete submachine in the homing FSM S^L_{homing}. Once $L = \Pi_n + n - 1$, there exists no complete submachine in the homing FSM S^L_{homing} and a corresponding HTC can be derived based on this homing FSM. The flow tables of homing FSMs S^5_{homing} and S^6_{homing} for the FSM S_1 in Figure 1 are shown in Figures 3 and 4, while a homing test case with minimum height derived using the FSM S^6_{homing} is shown in Figure 5. Due to Proposition 3, we apply input x_1 five times in order to reach the subset $\{2, 5\}$ that can be homed by x_2. Correspondingly, the height of a shortest homing test case becomes $6 = \Pi_2 + 2 - 2$.

Therefore, differently from non-initialized machines the problem of checking the existence and deriving an adaptive homing sequence for complete observable weakly initialized FSMs is not in **P**.

St/inputs	{1,3}	{2,4}	{1,5}	{4,5}	{2,4}/0	{2,3}	{1,4}	{2,5}	F
x_1	{2,4}/0	{1,5}/0	{2,3}/0	F/0,1,2	F/0,1,2	{1,4}/0	{2,5}/0	{1,3}/0	F/0,1,2
x_2	F/0,1,2	F/0,1,2	F/0,1,2	F/0,1,2	{2,4}/0	F/0,1,2	F/0,1,2	-	F/0,1,2

Fig. 4. The flow table of FSM S^5_{homing}

St/inputs	{1,3}	{2,4}	{1,5}	{4,5}	{2,4}/0	{2,3}	{1,4}	F
x_1	{2,4}/0	{1,5}/0	{2,3}/0	F/0,12	F/0,12	{1,4}/0	F/0,12	F/0,12
x_2	{1,3}/1	F/0,1,2	F/0,1,2	F/0,1,2	{2,4}/0	F/0,1,2	F/0,1,2	F/0,1,2

Fig. 5. The flow table of FSM S^6_{homing}

IV. Conclusions

The problem of checking the existence and deriving an adaptive homing / synchronizing sequence for complete non-initialized possibly nondeterministic observable FSMs is known to be in **P**. However, it is not the case for weakly initialized FSMs. In this paper, we have shown that given $n > 1$, there exists a complete weakly-initialized FSM S_n such that height of a shortest HTC is exponential with respect to the number of states / transitions of the FSM and thus, the problem of checking the existence and deriving an adaptive homing sequence is at least NP-hard. Correspondingly it is very important to determine classes of weakly initialized FSMs for which it is not the case. As it is illustrated by an FSM in Figure 1, it does not help to have a strongly connected FSM. However, it can be the case when the number of subsets reachable from the set of initial states is polynomial with respect to the number of states / transitions of the initial FSM. For example, it happens when the set of initial states is not large having three or four initial states, as it can well happen for real discrete event systems. We also notice that the results of this paper give a hint that the problem of checking the existence and deriving an adaptive synchronizing sequence is also at least NP-hard.

Acknowledgment

The work is supported by RFBR project No. 19-07-00327/19.

References

[1] 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)

[2] 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)

[3] T. S. Chow: Test design modeled by finite-state machines. IEEE Transactions on Software Engineering. 4(3): Pp. 178-187 (1978)

[4] Gill A. Introduction to the Theory of Finite-State Machines. 1964, 272 p.

[5] Kohavi, Z.: Switching and Finite Automata Theory. McGraw-Hill, New York (1978)

[6] 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)

[7] D. Lee, M. Yannakakis: Testing finite-state machines: state identification and verification. IEEE Trans. on Computers. 43(3): Pp. 306-320 (1994)

[8] Ana R. Cavalli, Caroline Gervy, Svetlana Prokopenko: New approaches for passive testing using an Extended Finite State Machine specification. Inf. Softw. Technol. 45(12), Pp. 837-852 (2003)

[9] 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)

[10] A. Petrenko, N. Yevtushenko: Adaptive Testing of Deterministic Implementations Specified by Nondeterministic FSMs. Lecture Notes in Computer Science (LNCS), № 7019, Pp. 162-178 (2011)

[11] N. Kushik. Methods for deriving homing and distinguishing experiments for nondeterministic FSMs. PhD thesis, Tomsk State University, 137 p.(2013)

[12] Kushik, N., El-Fakih, K., Yevtushenko, N., Cavalli, A. On adaptive experiments for nondeterministic finite state machines. International Journal on Software Tools for Technology Transfer. Vol. 18, N. 3, pp. 251-264 (2016)

[13] N. Kushik, N. Yevtushenko: Adaptive Homing is in P. Electronic Proceedings in Theoretical Computer Science, 180: Pp. 73-78 (2015)

[14] 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)

[15] A. Tvardovskii, N. Yevtushenko: Deriving adaptive distinguishing sequences for Finite State Machines. Trudy ISP RAN/Proc. ISP RAS, 30 (4), Pp. 139-154 (2018)

[16] 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)

[17] Vinarskii E., Tvardovskii A., Evtushenko L., Yevtushenko N. Deriving adaptive homing sequences for weakly initialized nondeterministic FSMs. Proceedings of IEEE East-West Design & Test Symposium, Pp. 1-6 (2019)

[18] N. Kushik, N. Yevtushenko. On the Length of Homing Sequences for Nondeterministic Finite State Machines. Lecture Notes in Computer Science, № 7982: Pp. 220-231 (2013)

[19] 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)

[20] Yevtushenko, N., Kuliamin, V.V., Kushik, N.: Evaluating the complexity of deriving adaptive homing, synchronizing and distinguishing sequences for nondeterministic fsms. In: Testing Software and Systems, 31st IFIP WG 6.1 International Conference, ICTSS 2019: Pp. 86–103 (2019)

[21] S. Sandberg: Homing and Synchronization Sequences. Model Based Testing of Reactive Systems, LNCS № 3472: Pp. 5-33 (2005)

[22] 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)

[23] 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)

[24] V. Alekseev: Introduction in theory of computation. (2002)

Deriving Distinguishing Sequences for Input/Output Automata

Igor Burdonov
Software Engineering department
Ivannikov Institute for System Programming of RAS
Moscow, Russia
igor@ispras.ru

Nina Yevtushenko
Software engineering deparment
Ivannikov Institute for System Programming of RAS
Moscow, Russia
evtushenko@ispras.ru

Alexander Kossachev
Software Engineering department
Ivannikov Institute for System Programming of RAS
Moscow, Russia
kos@ispras.ru

***Abstract* — Input/Output (I/O) automata are widely used when deriving high quality tests for (components of) complex discrete systems based on so called distinguishing sequences. For I/O automata, the number of distinguishability relations is much bigger than for classical deterministic Finite State Machines (FSM). In order to avoid submitting the same test sequence several number of times, i.e., avoid the "all weather conditions" assumption, the separability relation can be considered. If two Input/Output automata are separable then there exists an input sequence such that after submitting this sequence and observing produced outputs it can be uniquely concluded which automaton is under testing. In this paper, we modify the discipline of applying input sequences and discuss the derivation of separating sequences for automata with mixed states, i.e., states where transitions both under inputs and under outputs are defined, as well as with cycles labeled by outputs. We also illustrate how an adaptive separating sequence can be derived when both automata are input-enabled.**

Keywords—Finite State Machine (FSM), Input/Output (I/O) automata, distinguishing sequence

I. Introduction

When checking functional and non-functional requirements for complex programmable systems, a so-called mutation testing approach is widely used. The system specification is mutated using most crucial faults and an input sequence is derived that distinguishes the specification and its mutation if such a sequence exists, i.e., the distinguishability (separability) notion allows to formally define the conformance relation between the specification and an implementation under test. Distinguishing sequences are thoroughly studied for deterministic complete Finite State Machines (FSMs) [1] when at each state for each input there is exactly one transition. However, components of complex discrete systems can be partial and have the non-deterministic behavior. Moreover, the requirement to have an output directly after each input is too strong, since a system under test can response with an output (sequence) only to a sequence of inputs. Therefore, Input/Output (I/O) automata [2, 3] are more appropriate when describing the behavior of components of complex discrete systems. Nevertheless, the number of studies of distinguishing sequences for this model is much less.

For deterministic complete FSMs the distinguishability relation means that an implementation is not equivalent to the specification FSM that is there exists an input sequence such that the specification and implementation FSMs have different output responses to this sequence. For transition systems which have the nondeterministic behavior and are only partially specified, the number of distinguishability relations is much bigger. Moreover, in this case, a distinguishing sequence with respect to the nonequivalence or non-reduction relations has to be applied several times while its length can be exponential with respect the number of states of the specification if the specification is not observable [4, 5]. In order to avoid submitting the same test sequence several number of times the separability relation can be considered [6]. If two Input/Output automata are separable then there exists an input sequence such that after submitting this sequence and observing produced outputs it can be uniquely concluded which automaton is under experiment. In [7], such sequences are considered for I/O automata without mixed states, i.e., states where transitions both under inputs and under outputs are defined and there are no cycles labeled by outputs. In this paper, we extend the set of considered automata modifying the discipline of applying input sequences. We also illustrate how an adaptive separating sequence can be derived when both automata are input-enabled.

The rest of the paper is structured as follows. Section 2 contains the preliminaries. Section 3 illustrates how an (adaptive) separating sequence can be derived for a proper class of automata when at each state either input actions or output actions are specified. Section 4 proposes a method for deriving a separating sequence for general I/O automata.

II. Preliminaries

A finite Input/Output (I/O) automaton or simply an automaton throughout this paper is a 4-tuple $\boldsymbol{S} = (S, s_0, I, O, h_S)$ where S is a nonempty finite set of states with the designated initial state s_0, I and O are finite input and output alphabets, $I \cap O = \varnothing$, and $h_S \subseteq S \times (I \cup O) \times S$ is the transition relation. There exists a transition from state s to state s' under action a if and only if a triple $(s, a, s') \in h_S$. A state of the automaton is a *mixed* state if at the state, transitions under both inputs and outputs are

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

defined. The automaton is *observable*[1] if at each state under each action there exists at most one transition. The automaton is *nondeterministic* if at some state several output actions are specified [5]. In this paper, we consider only observable possibly nondeterministic automata if the converse is not directly stated. A trace of the automaton is a sequence of actions of $I \cup O$ that is permissible at the initial state. Given a trace σ, σ_{in} and σ_{out} are input and output projections of trace σ.

Denote S_{in} a subset of states where transitions under outputs are not specified. A trace at the initial state is *complete* if it is terminated at a state of the set S_{in}. In order to be able to observe such traces, a designated quiescence output $\delta \notin I \cup O$ is introduced [2, 3]; in other words, at each state where transitions under outputs are not specified a loop labeled by δ is added. The obtained automaton is denoted by $\boldsymbol{S}^{\delta}$ and by definition, δ is considered as an output. Therefore, a trace σ of $\boldsymbol{S}$ is complete if and only if $\boldsymbol{S}^{\delta}$ has a trace $\sigma\delta$, the latter corresponds to the fact that after this trace no output of the set O can be produced. Traces of $\boldsymbol{S}$ and $\boldsymbol{S}^{\delta}$ are closely related: given a trace of $\boldsymbol{S}^{\delta}$, after deleting δ a trace of the automaton $\boldsymbol{S}$ is obtained, and vice versa, given a trace σ of $\boldsymbol{S}$, if any number of δ are added after any prefix of σ that is complete then a trace of $\boldsymbol{S}^{\delta}$ is obtained.

III. Automata without mixed states and cycles with outputs

A. Preset separability

When using "white model" based testing there is a need to distinguish two automata by an experiment. Two automata $\boldsymbol{S}$ and $\boldsymbol{P}$ are separable then an automaton under experiment can be uniquely recognized after applying a separating input sequence α and observing a corresponding output sequence. If automata $\boldsymbol{S}$ and $\boldsymbol{P}$ have no mixed states and cycles labeled by outputs then in [7], a method is proposed how to check if two automata are separable and if yes how to derive a separating sequence under the *following hypothesis about applying input sequences* [9]. Before applying the first or the next input the tester waits an appropriate timeout T_{out}, i.e., the separating experiment with an automaton is organized in the following way. The tester waits for the timeout T_{out}, if a system under test produces an output then the timer advances from zero and the tester again waits until the timeout T_{out} expires. If during the timeout there is no output produced then the system is assumed to produce δ. After this, the tester applies the next input and waits again for T_{out}. Under this assumption, the separability problem can be solved for FSMs $\boldsymbol{M_S}$ and $\boldsymbol{M_P}$ which can be constructed for automata $\boldsymbol{S}$ and $\boldsymbol{P}$.

The set of $\boldsymbol{M_S}$ states is the set $S_{in} \cup \{s_0\}$; the initial state of $\boldsymbol{M_S}$ is s_0. FSM $\boldsymbol{M_S}$ is a 5-tuple (S_{in}, s_0, $I \cup \{null\_in\}$, $O \cup O^2 \cup \ldots \cup O^{ns} \cup \{\delta\}$, T_{MS}), $null\_in \notin I$, where ns is maximum length of a trace of $\boldsymbol{S}$ that has only outputs. The transition relation T_{MS} is the following. For each state $s \in S_{in}$ such that $(s, i, s') \in T_S$, $s' \in S_{in}$, T_{MS} has the transition (s, i, δ, s'), and for each state $s \in S_{in}$ such that $(s, i, s') \in T_S$, $s' \notin S_{in}$, T_{MS} has a transition $(s, i, o_1\, o_2 \ldots o_k, s'')$, $k \le ns$, where $s'' \in S_{in}$ is the $o_1\, o_2 \ldots o_k$-successor of state s'. If the initial state of $\boldsymbol{S}$ is not in S_{in}, then T_{MS} has a transition $(s_0, null\_in, o_1\, o_2 \ldots o_k, s)$, where $s \in S_{in}$, and s the $o_1\, o_2 \ldots o_k$-successor of state s_0.

If the automaton $\boldsymbol{S}$ ($\boldsymbol{P}$) is observable then the corresponding FSM $\boldsymbol{M_S}$ ($\boldsymbol{M_P}$) is also observable but can be partial and nondeterministic. In [7], it is shown that automata $\boldsymbol{S}$ are $\boldsymbol{P}$ are separable if and only if FSMs $\boldsymbol{M_S}$ and $\boldsymbol{M_P}$ are separable. For checking the FSM separability and deriving a separating sequence, a method from [10] can be utilized. Let $\boldsymbol{M_S}$ and $\boldsymbol{M_P}$ be separable with a separating sequence α. If α is headed by an input of I, then α is a separating sequence for automata $\boldsymbol{S}$ and $\boldsymbol{P}$. If $\alpha = null\_in\ \beta$ where $null\_in$ is a so-called empty input then β is a separating sequence for $\boldsymbol{S}$ and $\boldsymbol{P}$. If a separating sequence is applied to an automaton under experiment that is $\boldsymbol{S}$ or $\boldsymbol{P}$, then under the above hypothesis for applying input sequences it is possible to uniquely determine which automaton is under experiment.

Example 1. For automata $\boldsymbol{S}$ and $\boldsymbol{P}$ in Figs. 1 and 3 with initial states 1 and a, the corresponding FSMs $\boldsymbol{M_S}$ and $\boldsymbol{M_P}$ are in Figs. 2 and 4. These FSMs are not separable and thus, automata $\boldsymbol{S}$ and $\boldsymbol{P}$ also are not separable.

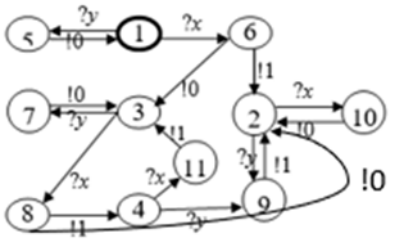


Fig. 1. Automaton $\boldsymbol{S}$

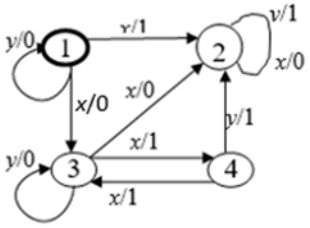


Fig. 2. FSM M_S

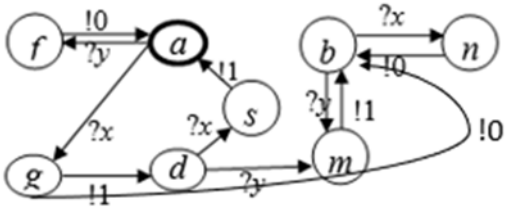


Fig. 3. Automaton $\boldsymbol{P}$

[1]Very often such an automaton is called deterministic [8]. However, we use the word «deterministic» for observable automata where at any state at most one output is specified.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

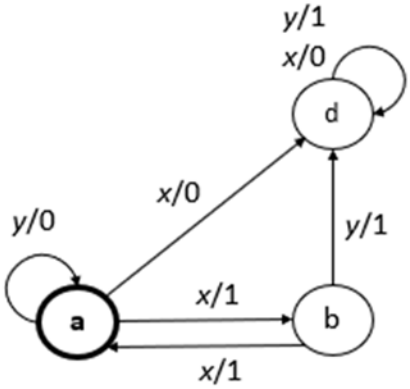


Fig. 4. FSM M_P

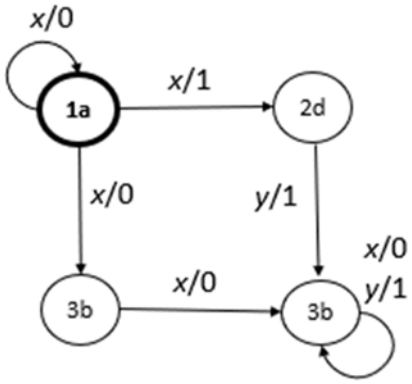


Fig. 5. The intersection of M_S и M_P

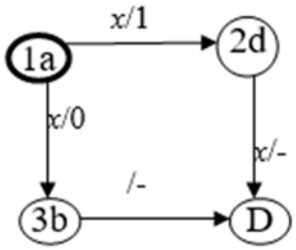


Fig. 6. A test case representation of an adaptive separating sequence

B. *Adaptive separability of automata without mixed states*

In the paper [13], it is shown that the length of a separating sequence when it exists can exponentially depend on the number of states at least of one automaton. In order to reduce such length, an adaptive separating experiment with automata can be considered where a separating sequence becomes adaptive, i.e., the next input significantly depends on the outputs to the previous inputs. An adaptive separating sequence can be represented by an acyclic automaton without mixed states: at a state where outputs are not defined there is at most one input while at each intermediate state where outputs are defined there is a transition under each output including δ.

An automaton $\boldsymbol{T} = (T, t_0, I, O \cup \{\delta\}, h_T)$ without mixed states that has an acyclic transition graph is a *test case* for automata defined over input alphabet I and output alphabet O if at each non-deadlock state, either at most a single input or all the outputs are defined and each complete trace is tailed by an output or δ.

Automata $\boldsymbol{S}$ and $\boldsymbol{P}$ without mixed states which are defined over input alphabet I and output alphabet O are *adaptively separable* if there exists a test case such that each complete trace is at most in one of these automata. Automata are *adaptively nonseparable* if for each test case there exists a complete trace that is a trace of both automata.

Since there is one-to-one correspondence between traces of automaton $\boldsymbol{S}^{\delta}$ ($\boldsymbol{P}^{\delta}$) and FSM M_S (M_P) [7], automata $\boldsymbol{S}$ and $\boldsymbol{P}$ are adaptively separable if and only if FSMs M_S and M_P are adaptively separable. If at each state of sets of S_{in} и P_{in} of observable automata $\boldsymbol{S}$ and $\boldsymbol{P}$ the behavior is defined under each input then FSMs M_S and M_P are complete and observable, and for checking their adaptive separability the following theorem can be used.

Theorem 1 [11, 12]. Observable complete initialized FSMs are adaptively separable if and only if their intersection has no complete submachines.

Example 1 (continuing). The intersection of FSMs M_S and M_P in Figures 2 and 4 is shown in Figure 5. Given a pair of states, there is an undefined transition in the intersection if FSMs at these states have no common outputs. A test case representing an adaptive separating sequence is shown in Figure 6.

IV. Separating automata with mixed states

Consider an automaton that has a state where both inputs and outputs are defined. In order to avoid races between inputs and outputs at such a state, another input timeout T_{in} is introduced. Until this timeout expires, the automaton expects an input. If there is no input before the timeout T_{in} expires then the automaton produces one of the prescribed outputs and moves to the next state or produces a quiescence output δ when the timeout T_{out} expires. Thus, by definition, T_{out} is always bigger than T_{in}. Given the timeout T_{in}, the timer starts to advance from zero after submitting an input or observing an output. When δ is produced, an input can be applied at any time instance.

In order to derive a separating sequence for automata with mixed states or with cycles labeled by output actions, we propose to transform such an automaton $\boldsymbol{S}$ into an automaton $\boldsymbol{S}^{\omega}$ without mixed states. For this purpose, we add a special input ω into $\boldsymbol{S}$: the input ω means that we need to wait T_{out} without submitting any input. Correspondingly, at each state s where outputs are defined, a transition to state s' is added under new input ω. All the transitions under outputs at s and only they are moved to state s'. Thus, the number of states in $\boldsymbol{S}^{\omega}$ is increased by the number of states where there are transitions under outputs, i.e., at most twice. An automaton $\boldsymbol{S}^{\omega}$ has no mixed states neither cycles labeled by output actions. For each state of $\boldsymbol{S}^{\omega}$ where are no transitions under outputs and ω, a loop labeled by ω is added to $\boldsymbol{S}^{\omega}$. At the next step, FSM M^{ω}_S is derived for the automaton $\boldsymbol{S}^{\omega,\delta}$ (Section 3a) with a small exception. If there is a loop at state of $\boldsymbol{S}^{\omega,\delta}$ labeled by ω и δ, then at this state a loop labeled by ω/δ is added to the FSM M^{ω}_S.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

Example 2. Consider an automaton $\boldsymbol{Q}$ in Figure 7 for which an automaton $\boldsymbol{Q}^{\omega}$ is shown in Figure 8 while the FSM M^{ω}_{Q} is in Figure 9.

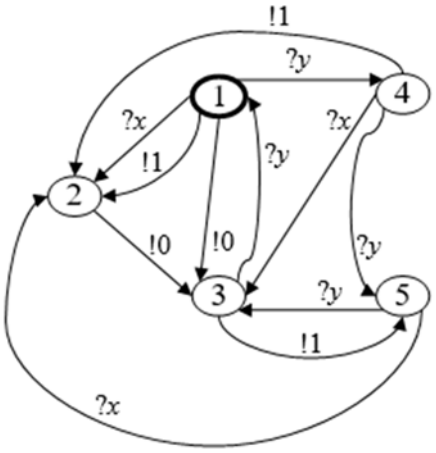


Fig. 7. Automaton $\boldsymbol{Q}$

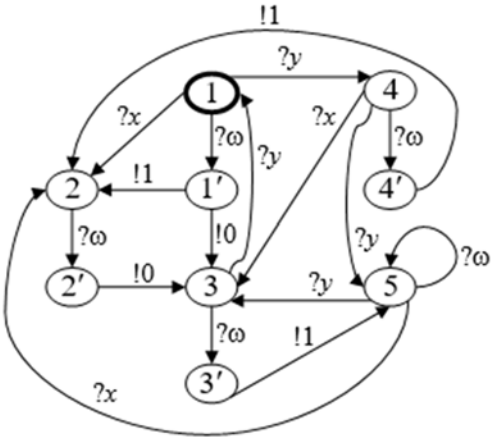


Fig. 8. Automaton $\boldsymbol{Q}^{\omega}$

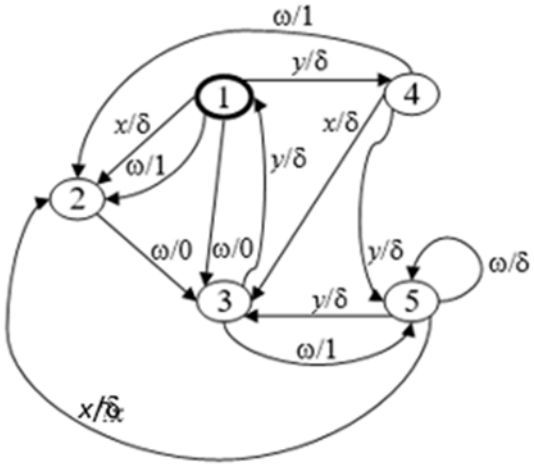


Fig. 9. FSM $M^{\omega}_{\boldsymbol{Q}}$

Automata $\boldsymbol{S}$ and $\boldsymbol{P}$ with mixed states are *separable* if automata $\boldsymbol{S}^{\omega}$ and $\boldsymbol{P}^{\omega}$ are separable. If automata $\boldsymbol{S}^{\omega}$ and $\boldsymbol{P}^{\omega}$ are separable then a *separating sequence* for these automata is a separating sequence for automata $\boldsymbol{S}$ and $\boldsymbol{P}$. However, the *hypothesis for applying input sequences* becomes more complex. An automaton under experiment waits for an input until the input timeout T_{in} expires and thus, a tester has to apply inputs fast enough, and a timer for calculating T_{in} advances from zero after applying a current input or receiving an output different from δ. After producing δ *under input* ω, in general, the next input can be applied at any time instance but for having unique requirements for applying inputs we assume that the output timeout T_{out} has to be over. When a separating sequence has an input ω, this means that no input is applied to an automaton under experiment: we just wait for an output to be produced during the output timeout T_{out} and only after an output is produced or T_{out} expires the next input can be applied (in the T_{in} range).

As an example, consider how a sequence *yx*ωω is applied to automaton $\boldsymbol{Q}$ (Fig. 7). An input *y* is applied and after this, the next input *x* is applied in the T_{in} range; thus, the automaton reaches state 3. After this an output is expected until the timeout T_{out} expires. In our example, the output !1 is produced. We wait another timeout T_{out} and obtain δ as an output. The experiment is over, since there are no more inputs in the input sequence. We cannot apply an input sequence *xy* at the initial state of $\boldsymbol{Q}$ since after accepting input *x* the automaton reaches state 2 where a transition under *y* is not defined. By construction of $\boldsymbol{S}^{\omega}$, the following statement holds.

Theorem 2. Given a trace σ^{ω} of $\boldsymbol{S}^{\omega\delta}$, a trace of $\boldsymbol{S}^{\delta}$ is obtained after deleting ω from the trace σ^{ω}, and vice versa, given a trace σ of $\boldsymbol{S}^{\delta}$, if ω is added in front of each output (except δ) and any number of ω in front of δ and after δ, then the obtained sequence is a trace of $\boldsymbol{S}^{\omega\delta}$.

Automata $\boldsymbol{S}^{\omega}$ and $\boldsymbol{P}^{\omega}$ have no mixed states and in order to check their separability and derive a separating sequence (if the automata are separable), FSMs M^{ω}_{S} and M^{ω}_{P} for automata $\boldsymbol{S}^{\omega,\delta}$ и $\boldsymbol{P}^{\omega,\delta}$ are derived. If FSMs M^{ω}_{S} and M^{ω}_{P} are separable then a separating sequence for these FSMs is a separating sequence for automata $\boldsymbol{S}$ and $\boldsymbol{P}$.

Indeed, if α is a separating sequence for automata $\boldsymbol{S}$ and $\boldsymbol{P}$ then α is a separating sequence for FSMs M^{ω}_{S} и M^{ω}_{P}. If α is a shortest separating sequence for FSMs M^{ω}_{S} и M^{ω}_{P}, then α = βω…ω and any proper prefix of α is not a separating sequence. Therefore, β takes both automata to states where the sets of output to ω…ω do not intersect, and thus, having a response to ω…ω we can conclude which automaton $\boldsymbol{S}$ or $\boldsymbol{P}$ is under experiment.

Given an automaton $\boldsymbol{S}$ in Figure 1, construct a corresponding automaton $\boldsymbol{S}^{\omega}$ and derive a corresponding FSM $M^{\omega}_{\boldsymbol{Q}}$ (Figure 9) for the automaton $\boldsymbol{Q}$ in Figure 7. FSMs $M^{\omega}_{\boldsymbol{Q}}$ и $M^{\omega}_{\boldsymbol{S}}$ are separated by an input sequence ω, since $\boldsymbol{S}$ after waiting for an output during T_{out} does not produce any output while $\boldsymbol{Q}$ can produce outputs !0 or !1.

V. Conclusions

In this paper, we have proposed a technique for separating Input/Output automata without a nonobservable action and in fact, this paper is the extension of [13] where separating

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

sequences are derived for I/O automata without mixed states, i.e., states where transitions both under inputs and under outputs are defined and there are no cycles labeled by outputs. In this paper, we extend the set of considered automata modifying the discipline of applying input sequences and the next step is to extend the obtained results for checking the existence and derivation of an adaptive separating sequence for Input/Output automata as well as to study other kinds of state identification sequences such as homing and synchronizing sequences.

ACKNOWLEDGMENT

This work is partly supported by RFBR project N 19-07-00327/19.

REFERENCES

[1] Kam, T., Villa, T., Brayton, K. R., Sangiovanni-Vincentelli, A. Synthesis of FSMs: Functional Optimization. Springer, 1997. 282 p.

[2] Tretmans J. A formal approach to conformance testing // The Intern. Workshop on Protocol Test Systems. 1993. P. 257–276.

[3] Burdonov I.B., Kossachev A.S., Kuliamin V.V. Conformance theory for systems with blocking and destruction. Nauka, 2008. 412 p.

[4] Iksoon Hwang, Nina Yevtushenko, Ana R. Cavalli: Tight bound on the length of distinguishing sequences for non-observable nondeterministic Finite-State Machines with a polynomial number of inputs and outputs. Inf. Process. Lett. 112(7). 2012. P. 298-301.

[5] Igor Burdonov, Alexandr Kossachev, Nina Yevtushenko, Alexey Demakovn. "Evaluating the Length of Distinguishing Sequences for Non-Deterministic Input/Output Automata", Proceedings of 2019 IEEE East-West Design & Test Symposium (EWDTS), pp. 445-449

[6] Starke P. Abstract Automata. American Elsevier, 1972. 419 p.

[7] I. Burdonov, N. Yevtushenko, A. Kossachev. Distingusihing transitions systems with nondeterministic behavior. In Proc. Of Russian conference «Scientific Service in Internet», 2019, P. 177-187 (in Russian).

[8] Hopcroft J.E., Motwani, R., and Ullman J.D.: Introduction to Automata Theory, Languages, and Computation. Addison-Wesley, second edition. 2001.

[9] Kushik N., Yevtushenko N., Burdonov I., Kossachev A. Synchronizing and Homing Experiments for Input/output Automata // System Informatics. 2017. No. 10. P. 1–10.

[10] Kushik N., Yevtushenko N., Cavalli A.R. On Testing against partial nondeterministic machines // Intern. Conf. on the Quality of information and Communications Technology. 2014. P. 230–233.

[11] Petrenko A., Yevtushenko N. Conformance Tests as Checking Experiments for Partial Nondeterministic FSM // Lecture Notes in Computer Science. 2005. V. 3997. P. 118–133.

[12] Yevtushenko N., Kushik N. Some state identification problems for nondeterministic Finite State Machines. CTT, 2018. 190 p. (in Russian)

[13] Burdonov I., Yevtushenko N., Kossachev A. Separating Input/Output automata with nondeterministic behavior // Russian Digital Libraries Journal. 2020. T. 23, Vol. 2 (in Russian).

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

On the Issue of Using Digital Radio Communications of the DMR Standard to Control the Train Traffic on Russian Railways

Alexander Nikitin,
DSc, professor at "Automation and Remote Control on Railways" Department, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
nikitin@crtc.spb.ru

Alexander Manakov,
DSc, professor at "Automation and Remote Control on Railways" Department, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
manakoff_2@mail.ru

Igor Kushpil,
PhD student at "Automation and Remote Control on Railways" Department, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
i_kushpil@mail.ru

Alexander Kostrominov,
DSc, professor at "Electrical Communications" Department, Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia
triak@grozon.spb.ru

Alexander Osminin,
DSc, professor, Vice-chairman of Joint Scientific Council of JSC Russian Railways
at@osminin.com

***Abstract* — Presented research is dedicated to the gradual transition from analogue types of railway radio communications to digital radio standards, mainly to the DMR standard. It was found that there is a lack of theoretical basis for this problem. The article set and solved the following tasks: determination of the volume of information transfers between trains and railway infrastructure, calculation of the information load arising in this case in the radio channel, determination of the required number of radio channels depending on the intensity of train traffic and the frequency of information transmissions. Moreover, a formalization of the methodology for calculating the radio range of a frequency band of 160 MHz, the dependences of the radio range on the level of reliability, power of base station, antennas installation heights and types of track complexity was carried out. The theoretical models and the graphs can be useful in the design and operation of railway traffic control systems, the principle of which is based on the use of a digital radio channel.**

Keywords — digital mobile radio, DMR, train control system, traffic control, railway radio system, information volume, information load, radio channel, radio range.

I. INTRODUCTION

The growing demand for the economic efficiency of rail transportation creates the prerequisites for introducing more efficient control systems into the transportation process, which are based on new principles of interval regulation of train traffic using a radio channel and satellite positioning.

The basis of such systems is the idea of constantly performing train traction calculation, both by train equipment and radio block center (RBC) equipment, indicating the coordinate of the track to which each train is allowed to move. The initial data for the calculation are: information on the length, speed, acceleration, direction of movement, current location (coordinate), serviceability and integrity of trains. Moreover, these data are determined by each train independently [1 – 7]. The availability of a reliable digital radio channel between trains and wayside infrastructure is a prerequisite for the implementation of such systems.

Today, all over the world, analogue types of railway radio communications are gradually being replaced by digital ones. The latter have a number of well-known advantages: the ability to transfer data, protect information, the ability to connect with radio equipment via a standard interface, etc.

In general case, a railway digital radio communication network consists of: distributed stationary base radio stations (BS), radio modems (as part of train equipment), antenna-feeder devices, and a digital data network. Each BS monitors all radio modems that are within its radio coverage area and maintains constant communication with them, Fig. 1.

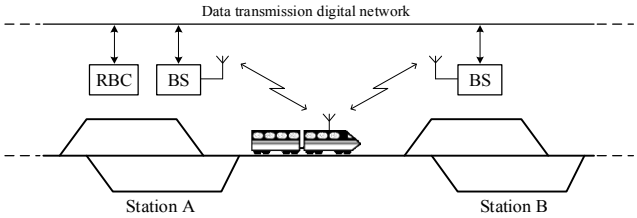


Fig. 1. The principle of organization of railway digital radio communications

Information from trains is sent to the BS and then to RBC, where the current train situation on the control area is determined and the coordinate is calculated, to which the safe movement of each train is guaranteed.

On the Russian railways, three frequency ranges have been reserved for the organization of digital radio data networks between trains and wayside infrastructure: 160 MHz, 460 MHz, 900 MHz. The GSM-R (900 MHz) and TETRA (460 MHz) digital radio standards widely known in the railway industry are quite expensive to implement, therefore they are provided only on high-speed lines (Moscow - St. Petersburg, Moscow - Adler, Moscow - Kazan, St. Petersburg - Helsinki).

The most flexible and less costly solution is to use the DMR radio standard. The DMR standard operates within the 12,5 kHz channel separation used worldwide in the land mobile radio frequency bands. TDMA technology allows getting two logical channels (bursts) from one physical channel. One burst consists of two information fields (each of 108 bits) and one 48-bit service field, Fig. 2. Data transfer, depending on the number of active bursts and the selected encoding method, is possible at speeds up to 9600 bps [8].

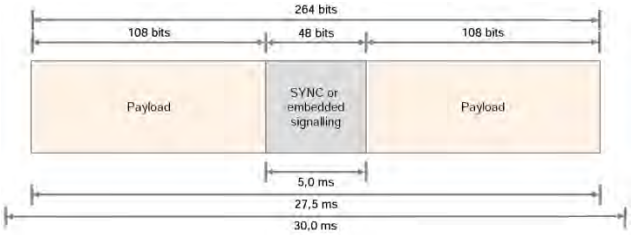


Fig. 2. Generic burst structure of DMR

Operating in the 160 MHz band, DMR equipment can be used to organize digital radio communications both at stations and on blocks. The DMR standard compares favorably with the GSM-R and TETRA standards due to the low cost of the equipment, its compatibility with equipment from various manufacturers and the ability to work simultaneously in two modes: analog and digital. This allows a flexible transition from analogue to digital radio communication systems with minimal costs.

At the first stage, it is proposed to preserve all types of existing analog radio communications, and to superimpose a digital DMR channel on them. Subsequently, a complete replacement of all analog equipment is expected. It is considered that soon the DMR standard will become widespread, mainly on trunk, secondary and low-density railway lines [9 – 11].

Nevertheless, the scientific investigations [1 – 11] showed that at present there is no theoretical basis necessary for the design and operation of digital radio networks of the DMR standard in railway transport. According to the authors of this article, the main tasks that need to be resolved in the first place are:

1. Development of a method that will allow estimating the volume of information transfers between on-board train devices and wayside equipment.
2. Obtaining a model that allows determining the information load created in a digital radio channel, the required number of radio channels and the probability of blocking data transmission, depending on the intensity of train traffic and the frequency of data transfers.
3. Formalization and adaptation a method for calculating the radio range (160 MHz) to the local area conditions, depending on the power of the BS, the height of the antennas and the operational features of stations and blocks, as well as obtaining dependency graphs.

II. DETERMINING THE VOLUME OF INFORMATION TRANSFERS

Two types of data are transmitted through a digital radio channel: voice and information. Thus, the task is reduced to determining the volume of one voice and information transfer.

In the DMR standard, one time slot (264 bits) is used to transmit 60 ms of voice [8]. Thus, 4400 bits are required to transmit 1 s of voice. To determine the approximate volume of information transfer to and from the train, the following approach is proposed.

Let $X=\{x_0, x_1, x_2,\ldots, x_n\}$ be the set of information packets transmitted to the train, where $\{x_0, x_1, x_2,\ldots, x_n\}$ are members of this set, with indices denoting the numbers of packets from the total number of all information packets transmitted to the train. At the same time, in the set X there is such a packet x_{const} (several packets) that can be conditionally considered constant, since it is present in each information transmission and contains the basic information necessary to provide continuous train movement. The other packets, such that $\forall x_{0\ldots n} \neq x_{const} \in X$, are conditionally variable and are transmitted as needed.

Let $Y=\{y_0, y_1, y_2,\ldots, y_n\}$ be the set of information packets transmitted from the train, where $\{y_0, y_1, y_2,\ldots, y_n\}$ are members of this set, with indices denoting the numbers of packets from the total number of all information packets transmitted from the train. With that, in the set Y there is such a packet y_{const} (several packets), which, by analogy with x_{const}, is considered conditionally constant. The other packets, such that $\forall y_{0\ldots n} \neq y_{const} \in Y$, are conditionally variable and are transmitted as needed.

Suppose, that in each information transfer to the train and back, there is only one conditionally constant and one conditionally variable information packet, then, the estimate of the volume of information transfer to the train $\overline{x}$ and from the train $\overline{y}$ will be determined as:

$$\begin{cases} \overline{x} = x_{const} + \sup\{x_{0\ldots n} \mid \neq x_{const} \in X\} \\ \overline{y} = y_{const} + \sup\{y_{0\ldots n} \mid \neq y_{const} \in Y\} \end{cases}, \quad (1)$$

where, sup $x_{0\ldots n}$ – exact upper bound (packet of the greatest volume, bit) of the set $X, \neq x_{const}$; sup $y_{0\ldots n}$ – exact upper bound of the set $Y, \neq y_{const}$.

III. INFORMATION LOAD IN DIGITAL RADIO CHANNEL

In the general case, for trunked radio systems with circuit switching (analog), the load in channel E [erlang] is determined according to Little's Law [12, 13]:

$$E = M\lambda h, \quad (2)$$

where, M – number of trains in the control area; λ – intensity of calls from one train in the busy hour; h – average time of occupation of the communication channel by one train in the busy hour, sec.

Model (2) as applied to trunking radio systems with packet switching (digital) has the form:

$$E = \frac{M\lambda L}{C}, \quad (3)$$

here, λ – intensity of receipt of data packets per unit time; L – average volume (length) of one packet, bit; C – bandwidth capability of the data transfer channel per unit time, bit/s.

Consider a model of the data transmission channel between the BS and the radio modems (RM), Fig. 3. Let $\lambda_1, \lambda_2,\ldots, \lambda_n$ – intensity incoming data packets with an average length l in RM1, RM2,..., RMn from BS per unit time; $\hat{\lambda}_1, \hat{\lambda}_2,\ldots, \hat{\lambda}_n$ – intensity incoming data packets with an average length $\hat{l}$ in BS from RM1, RM2,..., RMn per unit time.

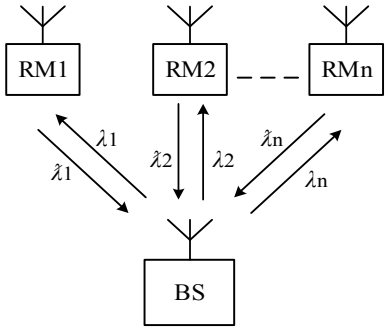


Fig. 3. Model of the data transmission channel between the BS and RM

Note, that $l \equiv \overline{x}$; $\hat{l} \equiv \overline{y}$; $l \neq \hat{l}$, see (1), which is correctly for the transmission of information data. When transmitting voice data, $l = \hat{l} = 4400$ bits, herewith $\lambda_1, \lambda_2,\ldots, \lambda_n$ and $\hat{\lambda}_1, \hat{\lambda}_2,\ldots, \hat{\lambda}_n$ are determined only by the duration of the conversation.

Thus, the total information load E arising in the DMR channel between BS and RM1, RM2,..., RMn is defined as:

$$E = \sum_{i=1}^{n} E_i = \frac{\sum_{i=1}^{n} \lambda_i l + \sum_{i=1}^{n} \hat{\lambda}_i \hat{l}}{C}. \quad (4)$$

To determine the capacity of the data transmission system (number of channels m) and the quality of service (probability of blocking the transmission P), various models are used in the teletraffic theory [14, 15]. The most widely used traffic models are Poisson (5) and Erlang B (6):

$$P(E,m) = 1 - e^{-E} \sum_{k=0}^{m-1} \frac{E^k}{k!}, \quad (5)$$

$$P(E,m) = \frac{\frac{E^m}{m!}}{\sum_{k=0}^{m} \frac{E^k}{k!}}. \quad (6)$$

They are based on the following assumptions:

- unlimited number of subscribers;
- random traffic flow;
- blocked transmission enters the queue (5);
- blocked transmission is redirected to other channels (6);
- exponential distribution of transmission hold time.

Given the requirements for the quality of service (P), the required number of radio channels (m) is determined by model (5) or (6). To simplify the calculations, special tables have been compiled [16]. It should be noted that the use of a particular model is determined by the features of the digital radio network.

IV. FORMALIZATION OF CALCULATION OF RADIO RANGE

To calculate the radio range of 160 MHz band at stations and blocks, it is proposed to adapt the method developed by VNIIZHT [17, 18]. To calculate, we introduce the following parameters:

r – radio range, km;

E' – field strength, dB;

U_2 – useful signal level at the input of the receiver (radio modem), dB;

$U_{2\,min}$ – level of the minimum allowable useful signal at the input of the receiver, dB. Depends on the type of traction: with diesel traction $U_{2\,min}$ = 4dB, with DC electric traction $U_{2\,min}$ = 6dB, with AC electric traction $U_{2\,min}$ = 14dB;

P_{bs} – signal power at the output of the base station, W;

α_1 (α_2) – attenuation in the coaxial cable of the transmitting (receiving) antenna, dB/m. Usually accepted: α_1=α_2=0.1dB/m;

l_1 (l_2) – length of coaxial cable of transmitting (receiving) antenna, m. Usually accepted: l_1=h_1, l_2=h_2;

B_a – signal attenuation coefficient, dB, due to signal attenuation in the coaxial cable of the transmitting (receiving) antenna, $B_a = \alpha_1 l_1 + \alpha_2 l_2$;

h_1 (h_2) – installation height of the transmitting (receiving) antenna, m. Usually h_2 is assumed to be 5 m (train height);

G_1 (G_2) – transmitting (receiving) antenna gains, dB. Usually accepted: G_1=G_2=0dB;

p – radio reliability, %. It is accepted from 97 - 99% for stations and from 93 - 95% for blocks, since at higher values, the veracity of the calculation results decreases.

To take into account local area conditions, equipment features, and signal interference phenomena, correction factors are introduced:

B_{dif1} – coefficient taking into account the difference of P_{bs} from 12W, dB, $B_{dif1} = 10\lg(P_{bs}/12)$;

B_{dif2} – coefficient taking into account the difference of P_{bs} from 1W, dB, $B_{dif2} = 10\lg(P_{bs}/1)$;

B_{loc} – signal attenuation coefficient by the locomotive body, dB. Usually accepted: B_{loc} = 9dB;

B_i – signal attenuation due to interference, dB. Depends on the required radio reliability p (if p = 97% B_i = -9dB; if p = 98% B_i = -11dB; if p = 99% B_i = -14dB), and applies only when calculating the station radio range;

K_{com} – track complexity coefficient from 1 to 5. As a rule, there is either a track of the second type (median terrain, K_{com}= 2), with height fluctuations of not more than 50m, or a track of the third type (light mountain terrain, K_{com} = 3);

K_{slow} – slow fluctuations in field strength due to changes in terrain, dB. For track type K_{com} = 2: if p = 93%, K_{slow} = 4.8dB; if p = 94%, K_{slow} = 5dB; if p = 95%, K_{slow} = 5.3dB. For track type K_{com} = 3: if p = 93%, K_{slow} = 6dB; if p = 94%, K_{slow} = 6.4dB; if p = 95%, K_{slow} = 6.7dB;

$\alpha_т$ – coefficient taking into account the peculiarities of radio wave propagation, dB, depending on K_{com} (if K_{com} = 2, $\alpha_т$ = 0dB; if K_{com} = 3, $\alpha_т$ = -3.4dB);

M – altitude coefficient, dB, taking into account the difference in the product of the installation heights of the antennas h_1 и h_2 from 100m^2, $M = 20\lg(h_1 h_2/100\text{m}^2)$;

K_{loc} – field attenuation coefficient of the metal roof of the locomotive, dB, depends on the type of locomotive and antenna. For example, with diesel traction on Russian railways, locomotives of the following types are used: М62, 2М62, ТЭМ2, 2ТЭМ116, ТЭМ18ДМ. For them, K_{loc} = 2 dB with a quarter-wave loop vibrator installed;

g_2 – coefficient of transition from signal field strength to voltage at the connection point of the receiving antenna with the feeder, dB. With a 75Ω feeder, g_2 = 10dB;

K_{int} – coefficient taking into account the presence of interference waves, dB. If p = 93% K_{int} = 1.8dB; if p = 94% K_{int} = 2dB; if p = 95% K_{int} = 2.2dB;

K_{flu} – fluctuations in field strength (daily and seasonal) due to changes in refraction in the troposphere, dB. If p = 93%, K_{flu} = 2.5dB; if p = 94%, K_{flu} = 2.8dB; if p = 95%, K_{flu} = 3.1dB;

K_{os} – field attenuation coefficient by overhead system, dB. For a single-track line, K_{os} = 1dB; for a double-track line, K_{os} = 2dB; in the absence of an overhead system, K_{os} = 0dB.

The calculation of the radio range (r) within the station is reduced to calculating the value of U_2, then, according to one of the approximating equations, the calculation r is performed [17]:

$$\begin{cases} U_2 = U_{2\,\min} + B_a - B_{dif1} + B_{loc} - B_i - G_1 - G_2 \\ r = \left(\left(-\dfrac{h_1 h_2}{160} \right) \ln\left(\dfrac{U_2 + 10}{110} \right) \right)^{0.5}, & \text{if } [90 \le U_2 < 100] \\ r = \left(\dfrac{77.5 h_1 h_2}{10^{\left(\frac{U_2+10}{20} \right)}} \right)^{0.5}, & \text{if } [20\lg(h_1 h_2) - 24 \le U_2 < 90] \\ r = \dfrac{-69 \ln\left(\dfrac{U_2 + 30}{93} \right)}{1 + \ln\left(\dfrac{1600}{h_1 h_2} \right)}, & \text{if } [-30 \le U_2 < 20\lg(h_1 h_2) - 24] \end{cases} \quad (7)$$

The calculation of the radio range (r) within the blocks is reduced to calculating the value of E', then, according to one of the approximating equations, the calculation r is performed [17]:

$$\begin{cases} E' = U_{2\min} - \alpha_т - B_{dif2} - G_1 - G_2 - M + B_a + K_{loc} + \\ + g_2 + K_{int} + K_{flu} + K_{slow} + K_{os} \\ r = 10 \exp\left(-\dfrac{(E' - 15)}{20.2} \right), & \text{if } [h_1 h_2 = 25\text{m}^2] \\ r = 10 \exp\left(-\dfrac{(E' - 25)}{19.6} \right), & \text{if } [h_1 h_2 = 100\text{m}^2] \end{cases} \quad (8)$$

Calculations of the radio range at stations and blocks at diesel traction, various BS power values, radio reliability level, antenna installation height and two types of track complexity, by models (7), (8), were carried out (see Fig. 4, 5, 6), [19].

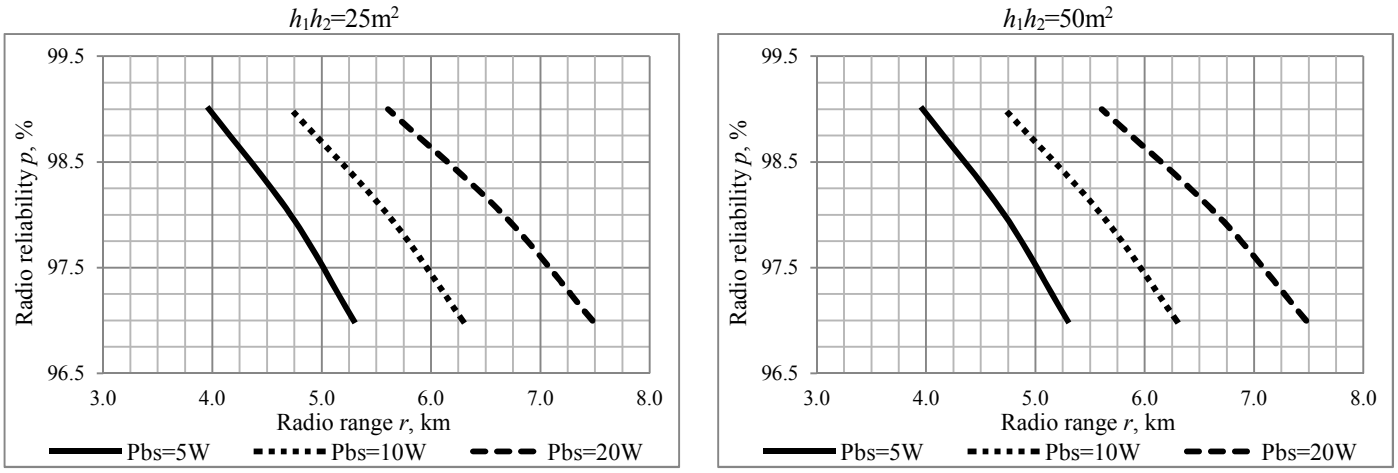


Fig. 4. Dependents graphs of the radio range at the station

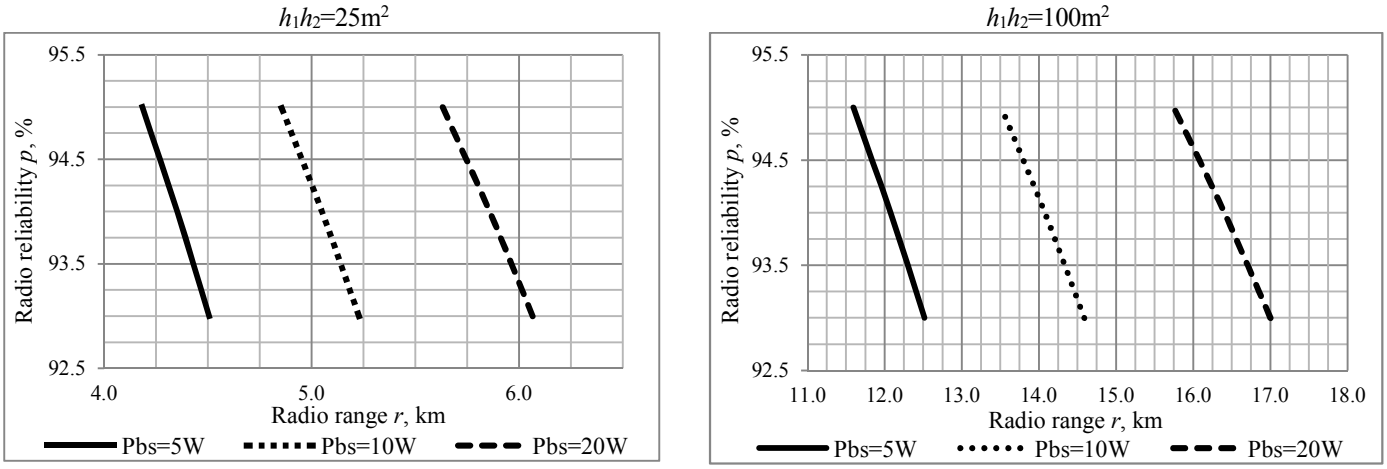


Fig. 5. Dependents graphs of the radio range at the blocks on the median terrain (K_{com} = 2)

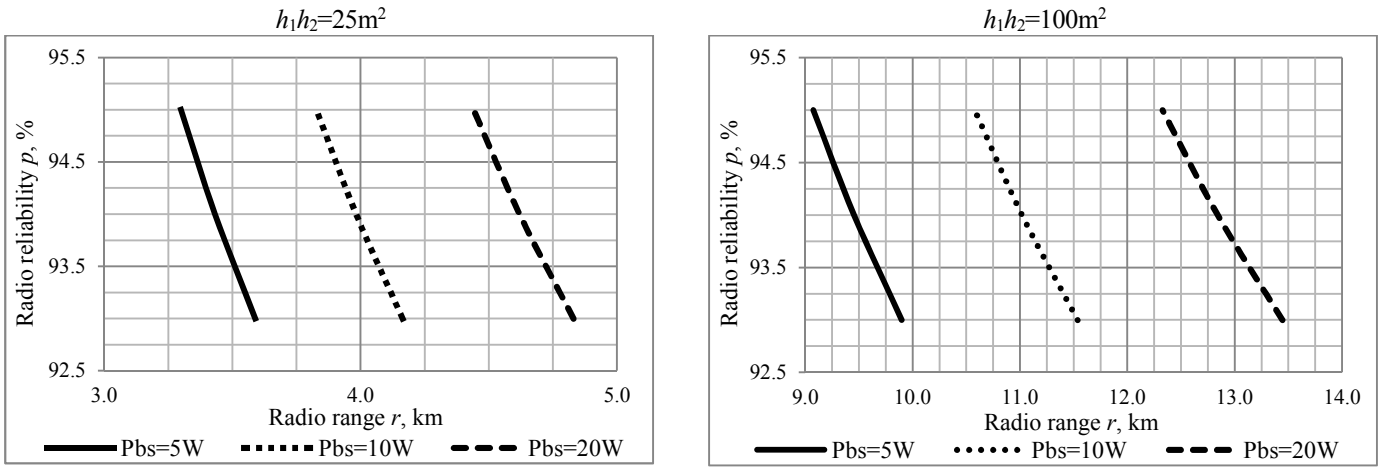


Fig. 6. Dependents graphs of the radio range at the blocks on the light mountain terrain (K_{com} = 3)

V. CONCLUSION

The use of digital radio communications in the railway transportation process is one of the main conditions for the introduction of more efficient train control systems. At the same time, the task of minimizing capital investments in digital radio communication systems, by preserving existing analogue communication systems, is becoming especially relevant. The research showed a growing demand for DMR radio communications, which compares favorably with the GSM-R and TETRA standards, as it allows making transition from analogue to digital communication systems at minimal cost.

The scientific novelty of the paper consists in the formulation and solution of tasks that arise when designing train control systems based on a digital radio channel, when it is necessary to take into account the radio range and the information load created in the radio channels. Mathematical models (1-8), as well as dependency graphs (Fig. 4, 5, 6) will find their application in the design and further operation of such systems.

The obtained graphs show that for the organization of station radio communication, one radio modem with a power of 5 W is sufficient, with a transmitting antenna installation height of 5 m. At the same time, a radio range of 4 km is provided with a radio reliability of 99%. This is sufficient, since the length of stations is rarely more than 2-3 km. If necessary, the range of radio communication can be increased.

For the organization of radio communication on the blocks, it is necessary to increase the transmitter power, the installation height of the transmitting antenna, or both of these parameters simultaneously. It is also necessary to take into account the type of terrain and length of the blocks. In addition, to improve the reliability of radio communication, it is desirable to use radio equipment that supports several frequency bands simultaneously, for example 160, 330, 450 MHz.

According to the authors, further promising research ways lie in the field of improvement and development of on-board train integrity systems, the principle of which is based on the use of digital radio communication of the DMR standard.

REFERENCES

[1] ERTMS/ETCS – Baseline 3. System Requirements Specification. Chapter 7. ISSUE: 3.0.0 DATE: 23/12/08.

[2] H. Scholten, R. Westenberg and M. Schoemaker, "Sensing Train Integrity," IEEE SENSORS 2009 Conference, pp. 669-674, 2009.

[3] Train Integrity Concept and Functional Requirements Specifications, 3 ed., 2018.

[4] A. Nikitin, N. Shatalova and I. Kushpil, "Materials of the Russian Scientific and Practical Conference "Transport of Russia: Problems and Prospects - 2019"," in Principle of construction and functioning algorithms of on-board train integrity control systems, Saint-Petersburg, 2019.

[5] I. Sassi and E. El-Koursi, "Proceedings of the 29th European Safety and Reliability Conference (ESREL)," in On-Board Train Integrity: Safety Requirements Analysis, 2019.

[6] H. Shigeto, F. Mitsuyoshi, F. Hiroyuki and O. Yuto, "Train control system for secondary lines using radio communications in specific area," Quarterly Report of RTRI, Vol. 53, no. 1, pp. 1-6, February 2012.

[7] J. Hyunjeong, K. Gonyop, B. Jonghyen, L. Kangmi and K. Yongkyu, "Development of the on-board centered train control system to enhance efficiency of low-density railway line," Computer Applications for Security, Control and System Engineering, pp. 269-276, 2012.

[8] ETSI TR 102 398 V1.1.1 (2006-05). Electromagnetic compatibility and Radio spectrum Matters (ERM); Digital Mobile Radio (DMR) General System Design.

[9] Order on approval of the generalized frequency plan of Russian Railways in the 160 MHz band No. 340r dated 02/11/2013..

[10] J. WANG, M. Cheng, C. Baigen and L. Jiang, "A train control system for low density line in China," Journal of the China railway society, pp. 46-52, December 2015.

[11] A. Nikitin, J. Kokurin, I. Kushpil and V. Sharov, "New method of trains movement for low-density railway lines of Russian Railways," Automation on Transport, no. 3, pp. 561-579, 2018.

[12] MPT 1318 Engineering Memorandum Trunked Systems in the Land Mobile Radio Service. Revised and reprinted January 1994.

[13] "Little's law," [Online]. Available: https://en.wikipedia.org/wiki/Little%27s_law. [Accessed 10 02 2020].

[14] "Traffic analysis," [Online]. Available: https://www.cisco.com/c/ru_ru/td/docs/ios/solutions_docs/voip_solutions/TA_ISD.html. [Accessed 07 02 2020].

[15] E. Chromy, J. Suran, M. Kovacik and M. Kavacky, "Usage of Erlang Formula in IP Networks," Communications and Network, pp. 161-167, 03 2011.

[16] "Traffic table," [Online]. Available: https://www.pitt.edu/~dtipper/erlang-table.pdf. [Accessed 09 02 2020].

[17] G. Gorelov, D. Roenkov and Y. Yurkin, Communication systems with moving objects, Moscow: Federal State Budget Educational Establishment "Educational and Methodological Center for Education in Railway Transport", 2014, p. 335.

[18] Metodicheskiye ukazaniya po organizatsii i raschetu setey poyezdnoy radiosvyazi OAO "RZHD" utverzhdeny rasporyazheniyem 2854r ot 23.12.2013g.

[19] A. Nikitin, I. Kushpil, "Investigation of the possibility of introduction of digital radio communications and organization of data transfer between stations at low-density lines" Automation on Transport, Vol. 5, no. 1, pp. 45-61, March 2019

Thermoregulation of smart clothing based on Peltier elements

Mikhail F. Mitsik
Mathematics and applied informatics
Don State Technical University,
Rostov-on-Don, Russia
m_mits@mail.ru

Marina V. Byrdina
Designing, technology and design
Don State Technical University
Rostov-on-Don, Russia
byrdinamarina@mail.ru

***Abstract*—The paper proposes a method of clothing creation that provides cooling (or heating) of the body based on a system of series-connected Peltier elements. The calculation of the power of heat loss for cooling the body as applied to smart clothing for the upper body, as well as the calculation of the power for cooling produced by the system of series-connected Peltier elements, is presented. It is shown that this design provides the designed cooling of the upper body by 5 K. Experimental studies of a system of series-connected Peltier elements were carried out and the effect of cooling one side of the surface to the required temperature was revealed.. Smart fabric is modeled as a flexible, inextensible multilayer shell. The dependence of the heat loss on cooling the body from the power generated by Peltier elements is nonlinear due to the fact that part of the electric current energy is spent on the thermal resistance of the electric circuit itself, which complicates the application of the proposed approach, but does not exclude the possibility of minimizing this effect in the future.**

Keywords— smart clothing, Peltier thermoelectric element, thin flexible inextensible shell, calculation of heat loss for cooling, calculation of thermoelectric element power.

I. INTRODUCTION

In recent years, the volume of industrial and household equipment that is created with cooling or thermostating using Peltier elements has been increasing (Peltie elements, Thermoelectric Cooler, TEC). Thermoelectric cooling elements, the so-called Peltier elements, are increasingly used for cooling various small-sized elements in electronic devices, such as microprocessors, CCD arrays and CCDs. Also, Peltier elements are used to stabilize the temperature of coherent sources of optical radiation in order to avoid their drift, and in a number of other applications [1]. Their important practical advantages include small dimensions and weight, the absence of coolants, the ability to cool the device significantly below the ambient temperature.

However, despite the improvement in the quality of developments based on Peltier elements and the production of more advanced elements, the issue of modern and reliable controllers for controlling Peltier elements in the proposed design is still relevant.

For a person the thermoregulation of the body is one of the main conditions for maintaining comfortable condition and life support. At the same time, a person can work and exist in a fairly narrow temperature range. To protect the body from hypothermia from the external environment, one of the most effective ways to keep the body warm is multilayer clothing. Multi-layer clothing allows keeping heat close to the body, on the other hand, air layers allow for breathability [1]. The main assistant to the body in this case may be layers of natural fabrics, in particular wool.

However, the temperature of the environment is subject to fluctuations, changes in air humidity and weather. A person who is out of the house may feel uncomfortable due to changing temperature conditions [2, 3]. Under such conditions, the body may be subject to both hypothermia and overheating.

One of the ways to maintain a comfortable temperature for life can be a system of temperature controllers based on Peltier cooling elements woven into a smart fabric [4, 5].

The purpose of the paper is to develop a way to create clothing that provides active thermoregulation of the body in the event of adverse environmental influences based on the Peltier temperature regulator.

Research objectives:

1) describe the principle of operation of Peltier temperature controllers as a set of thermocouples - semiconductor pairs, one of which is n-type and the other is p-type, a description of the operation of the thermocouple;

2) calculate the power supplied to the input of a system of Peltier elements woven into smart clothing, as well as the power of a system of Peltier elements providing a decrease in temperature on the inner surface of the upper part of clothing by 5 K.

The novelty of the research lies in the design of smart clothing with active thermoregulation based on a system of series-connected Peltier elements and the calculation of the heat loss power for cooling (or heating) the body.

It should be noted that with a large selection of various options for the execution of Peltier elements and assemblies based on them, at present there are few ready-made and high-quality devices for controlling the power supplied to the elements and, as a result, the temperature of the objects of cooling or heating themselves. These circumstances are associated with the fact that when low-frequency pulse-width modulation is applied to the Peltier elements, the efficiency of the elements is significantly reduced (up to 30%), and therefore, it is required to switch to the use of more advanced and, accordingly, more expensive power control schemes.

II. Description of the Peltier element and its characteristics

The Peltier element is a thermoelectric module consisting of a large number of thermocouples that are interconnected by switching plates, the plates are arranged in the form of rectangles [6, 7]. A thermocouple module is placed between a pair of thin ceramic plates (Fig. 1).

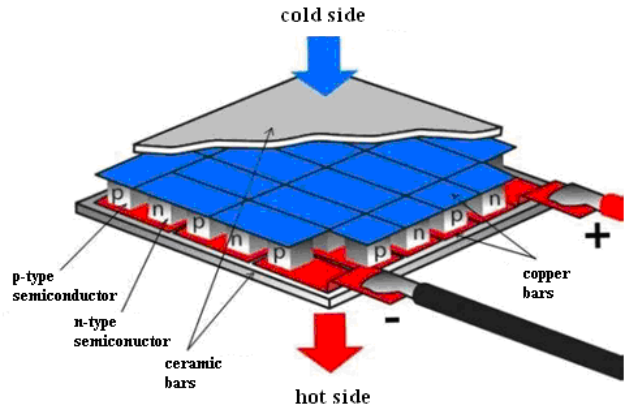


Fig. 1 – Peltier thermoelectric module design

The number of thermocouples may vary depending on the design and count several hundred, which will reduce the thickness of the thermoelectric module from a few millimeters to fractions of a millimeter. At the same time, the module power can reach several tens of watts [8].

The design of the Peltier element consists of several thermocouples [1, 9, 10], i.e. pairs of parallelepipeds built on semiconductors, one of which is n-type, and the other is p-type and which are interconnected by a metal jumper (fig. 2)

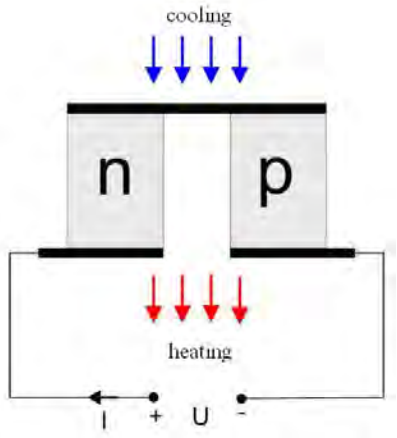


Fig. 2 – Schematic diagram of a thermocouple

Metal jumpers are contacts for heat transfer, on the other side they are insulated with a ceramic plate. Thermocouples are connected in such way that a series connection of a large number of pairs occurs so that on the upper surface there is one set of compounds (n-> p), and on the bottom - the opposite (p-> n). Electric current moves sequentially through a set of parallelepipeds, while the direction of current flow affects which of the surfaces: the top or bottom is heated, respectively, the opposite surface is cooled.

Under the influence of electric current, the Peltier element transfers heat from one surface of the thermoelectric module to the opposite and a temperature difference occurs. The amount of heat generated by the Peltier element, as follows from numerous experimental studies [10], can be described by the formula:

$$Q_P = P \cdot I \cdot t, \qquad (1)$$

where P – the so-called Peltier coefficient,

I – current flowing through an element,

t – current flow time.

As experiments [9] show, the Peltier coefficient substantially depends on temperature and a pair of dissimilar metals forming a thermocouple. The values of the Peltier coefficient at various temperatures and metal pairs are shown in table 1.

TABLE I. Peltier coefficient values versus temperature for metal pairs

Iron-constantan		**Copper-nickel**		**Lead-Constantan**	
T, K	P, mV	T, K	P, mV	T, K	P, mV
273	13,0	292	8,0	293	8,7
299	15,0	328	9,0	383	11,8
403	19,0	478	10,3	508	16,0
513	26,0	563	8,6	578	18,7
593	34,0	613	8,0	633	20,6
833	52,0	718	10,0	713	23,4

For the Peltier coefficient, which is one of the important thermoelectric characteristics of materials, there is a formula by which it is determined through the Thomson coefficient:

$$P = a \cdot T, \qquad (2)$$

where a – Thomson coefficient, T – absolute temperature.

It should be noted that the Thomson coefficient depends on the pair of metals on the basis of which the Peltier element is constructed.

Peltier elements, with regard to clothing, have both advantages and disadvantages. The advantages of using Peltier elements are the rather convenient and compact sizes of the elements for weaving them into the fabric, the possibility of their future creation on flexible boards, the absence of any moving parts, gases, liquids. When changing the direction of current flow through the element, it is

possible to change the cooling to heating, i.e. it is possible to solve the problem of thermoregulation. The disadvantages of modern Peltier elements are the relatively low efficiency and high power consumption of electricity. If you use the Peltier element for cooling, then heat will need to be removed from the hot surface. If this is not done, the cooling efficiency will be significantly reduced.

III. Calculation of the power of the cooling device of smart clothes

We calculate the necessary power to supply the Peltier elements located in the human smart clothing. For now, we will exclude the need for a significant change in the temperature of the surface of the human body; let heating or cooling the surface of the body to 5 K be required. Assuming that the working surface of the body is a cylinder of height $H = 0{,}5m$ and diameter $D = 0{,}3m$, we find that the surface area is

$$S = H \cdot D \cdot \pi = 0{,}15 \cdot \pi \left(m^2\right).$$

The overall dimensions of the Peltier element TEC1-12706 are $0{,}04 \times 0{,}04 \times 0{,}0039\left(m^3\right)$. The working surface of one Peltier element is equal to

$$S_P = 0{,}04 \times 0{,}04 = 0{,}0016\left(m^2\right)$$

Given the technological windows between the Peltier elements, we choose the dimensions for placing one element equal $0{,}06 \times 0{,}06\left(m^2\right)$

$$S_r = 0{,}06 \times 0{,}06 = 0{,}0036\left(m^2\right)$$

Approximately n units must be placed on the work surface of the garment

$$n = \frac{S}{S_r} = \frac{0{,}15\pi}{0{,}0036} = 131\left(units\right). \qquad (3)$$

Given the areas for the sleeves, on the chest and back of the clothing enough to place 120 units of Peltier.

Heat loss for cooling the body will be determined by the formula

$$Q = \frac{\lambda \cdot S \cdot dT}{h}, \qquad (4)$$

where: $\lambda, \frac{W}{m \cdot K}$ – thermal conductivity of smart fabric,

$dT = 5K$ – temperature difference on cold and hot surfaces of Peltier elements,

h – thickness of smart fabric, taking into account the air gap.

As experimental studies show, the average thickness of the air gap is

$$h_{Gap} = 0{,}002m.$$

The average thickness of smart fabric is

$$h_{Sm} = 0{,}001м.$$

Determine the thickness of the smart fabric, taking into account the air gap

$$h = h_{Gap} + h_{Sm} = 0{,}003м.$$

The average thermal conductivity of smart fabric is

$$\lambda_{Sm} = 0{,}07\frac{W}{m \cdot K}.$$

The coefficient of thermal conductivity of the air gap at body temperature is determined as tabular

$$\lambda_{Gap} = 0{,}027\frac{W}{m \cdot K}.$$

Thus, the total coefficient of thermal conductivity from Peltier elements to the body is

$$\lambda = \frac{\lambda_{Sm} \cdot h_{Sm} + \lambda_{Gap} \cdot h_{Gap}}{h} = 0{,}04133\frac{W}{m \cdot K}.$$

We find the heat loss for cooling the body according to the formula (4)

$$Q = \frac{0{,}04133 \cdot 0{,}471 \cdot 5}{0{,}003} = 32{,}5W. \quad (5)$$

We determine the performance of the Peltier elements. Let the operating voltage of 36 V be supplied to the input. The rated current for the Peltier element TEC1 12706 is $I_1 = 4A$. Let the resistance to the movement of electric current caused by one Peltier element be equal R_1. Ohm's law for a chain section

$$R_1 = \frac{U}{I_1} = 9\left(\Omega\right).$$

The resistance of the entire chain of 120 Peltier elements will be 120 times greater

$$R = 120 \cdot R_1 = 1080\left(\Omega\right).$$

Then the current strength in the entire circuit is also determined by Ohm's law

$$I = \frac{U}{R} = \frac{36}{1080} = \frac{1}{30}A.$$

The power generated by one Peltier element is equal to

$$Q_1 = U \cdot I = \frac{36}{30} = 1{,}2\ W. \qquad (6)$$

Expression (6) describes the theoretical value of the power of the element, in practice it is about 30% of the theoretical power

$$Q_R = Q_1 \cdot 0{,}3 = 0{,}36W.$$

The power of 120 Peltier elements will accordingly be equal

$$Q_S = Q_R \cdot 120 = 0{,}36 \cdot 120 = 43{,}2W. \quad (7)$$

Since the heat loss for cooling the body is less than the total power of the Peltier elements, i.e.

$$Q = 32{,}5\ W < Q_S = 43{,}2W,$$

then the proposed clothing provides 5 K body cooling, i.e. solves the problem.

The operating voltage can be provided either from a household AC network using a step-down transformer, or from a DC battery circuit.

It should be noted that the dependence of the heat loss on cooling the body on the power generated by Peltier elements is nonlinear due to the fact that part of the electric current energy is spent on the thermal resistance of the electric circuit itself, which leads to some heating of all clothing. In addition, for the work of the proposed clothing to cool effectively, the external surfaces of the Peltier elements that work for heating must be cooled, for example, by an external air stream. Otherwise, the efficiency of the structure may noticeably decrease.

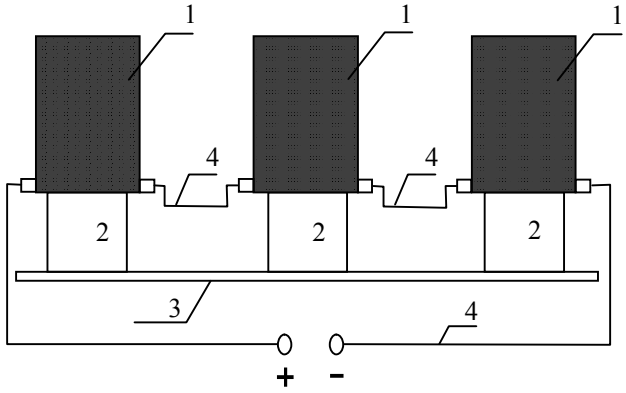
The main disadvantage the Peltier elements, despite the progress in the field of semiconductors, is the lower efficiency of the elements in comparison, for example, with compression cooling which is especially noticeable at high cooling capacities. However, in the case of using Peltier elements for the thermoregulation of clothing, there is no need for large cooling capacities; a temperature drop of 5-10 ° C in a small volume is quite enough for them. Under such conditions, a design based on Peltier elements becomes one of the promising solutions for generating cold and heat.

IV. Experimental studies of cooling clothing device

The number of thermocouples can be changed over a wide range: from units to hundreds, which makes it possible to design a thermoelectric modules with a refrigerating capacity from tenths to hundreds of watts with an operating voltage from tenths to tens of volts.

It should be noted that the temperature difference between the hot and cold sides of the thermoelectric module in the problem under consideration may vary from 5 ° C to 20 ° C. In this case, the greater the temperature difference, the less energy per unit time (in the form of heat) is transferred by the thermoelement. In the problem under consideration, the temperature difference between the hot and cold sides of the thermoelement takes on rather small values, i.e. significantly less than 70 ° C, which ensures efficient heat transfer by the thermoelectric module.

The experiment was carried out in a laboratory at a temperature of 27°C and minimal exposure to air currents. During the experiment, measurements of current-voltage characteristics and temperatures of colder and hotter surfaces of Peltier elements were performed. At the stage of the first experiment, three Peltier elements were connected in series in a circuit. To minimize heat flux through the table and fastening elements, each of the Peltier elements 1 was bolted to a thin (0.5 mm) aluminum plate 2 in air, so that the desktop 3 only touched the lower edge of the plate (Fig. 3).



1 – Peltier elements; 2 – aluminum plates;
3 – desktop cover; 4 – connecting wires.

Fig. 3 – Peltier element circuit testing machine

Since, according to the calculation, 120 Peltier elements should be contained in the circuit of the cooling device of smart clothes, and so far only three elements are involved in the experiment, the voltage applied to the input of the circuit of three elements will be 40 times less.

During the experiment, a voltage was applied to the input of the circuit with a step of 0.5 V, and measurements of the current strength and temperatures established on the cold and hot surfaces of the elements were carried out. The temperatures on each of the surfaces of the Peltier elements were measured with a Testo 0560 1110 mini-thermometer. The measurement results are entered in the table 2.

TABLE II. Results of measurements of current-voltage characteristics and temperatures on the circuit of Peltier elements

experiment No.	1	2	3	4
Cold side temperature, °C	22	20	18	15
Hot side temperature, °C	31	33	35	37
Temperature difference, °C	9	13	17	22
Amperage, A	0,3	0,5	0,7	1
Voltage, V	0,5	1	1,5	2

It should be noted that on each element the surface temperatures of the hotter sides practically did not differ, as well as the temperatures of the surfaces of the colder sides, respectively, in Table 2, the surface temperatures averaged for all three elements are presented. As it is easy to see from Table 2, the problem of thermoelectric cooling can be realized under experimental conditions already for a voltage of 0.5 V, or for a voltage slightly higher.

Since for smart clothes the initial temperature of the cold side is 36°C, then during the experiment the temperature of the hot side will increase in comparison with the values from Table 2, taking into account the temperature difference between the cold and hot sides. In this case, the temperature of the cold side should only drop to a temperature that is comfortable for humans. Thus, the voltage applied to the thermoelectric circuit must be adjustable for the convenience of the user.

V. Conclusions

The problem of active cooling or heating of smart clothing is very relevant in the modern world due to changes in temperature in the external environment. Overheating of the body, or its hypothermia entail a decrease in the working capacity and comfort of a person in a hot or cold environment. As an active approach to the body's thermoregulation, we propose smart clothing designed on the basis of the Peltier system of thermoelectric elements. In this work, we calculated the power of smart clothing for heat loss during cooling of the body, and also calculated the necessary power to power Peltier elements located in human smart clothing. Performed experimental studies have shown the technical possibility of achieving the effect of cooling the body using a series circuit of thermoelectric elements to the required temperature, if efficient cooling is possible for the hot side.

The undoubted advantages of using thermoelectric elements are their silent operation, the absence of moving parts and environments, compact dimensions, and the possibility of their integration into the clothing of the future. A change in the direction of current flow through the system of thermoelectric elements entails a change in the mode of operation of the system from cooling to heating, and thus the task of thermoregulation of the body is solved. Regulation of the power supplied to the power of thermocouples will regulate the power of heat loss for cooling the body. The results will be used to create packages of smart clothing in the future.

Despite numerous experimental studies illustrating the Peltier effect, in practice, in the development of cooling systems, the Peltier effect has not yet received enough attention. This is largely due to the lack of awareness of the developers of electronic devices about the main features of thermoelectric cooling systems, as a result of which, as a rule, an element is selected that is closest in terms of nominal power supply parameters to those already used in an electronic device.

At the same time, the maximum efficiency of the devices is not guaranteed per unit of electrical energy spent on cooling. The present study was undertaken in order to optimize the thermoregulation of clothing under conditions of limited permissible power consumption and to develop related technical solutions to achieve a comfortable state of the human body.

References

[1] H. Julian Goldsmid, Bismuth Telluride and Its Alloys as Materials for Thermoelectric Generation, Materials 2014, 7, 2577-2592.

[2] R. Chein, Y. Chen, “Performances of thermoelectric cooler integrated with microchannel heat sinks”, International Journal of Refrigeration 28 (2005) 828–839.

[3] S. Riffat, X. ma, Improving the coefficient of performance of thermoelectric cooling systems, international journal of energy research Int. J. Energy Res. 2004; 28:753–768.

[4] J. Vian, D. Astrain, “Development of a heat exchanger for the cold side of a thermoelectric module”, Applied Thermal Engineering 28 (2008) 1514–1521.

[5] Marc H., Thermoelectric Modules: Principles and Research, InterPACK july 6-8, 2011, Portland.

[6] Kaseb S., El-hairy G, Electronics Cooling, Mechanical Power Engineering Department, Faculty of Engineering, Cairo University, Egypt, 2014.

[7] Enescu D, Virjoghe EO, A review on thermoelectric cooling parameters and performance, Renewable and Sustainable Energy Reviews , 2014, 38:903–916.

[8] D. Zhao, G. Tan, A review of thermoelectric cooling: materials, modeling and applications, Applied Thermal Engineering (2014), doi: 10.1016/ j.applthermaleng. 2014.01.074.

[9] Andreas Larsson, Torleif A. Tollefsen, Ole Martin Løvvik, Knut E Aasmundtveit. Thermoelectric module for high temperature application. Conference: 2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm). DOI: 10.1109/ITHERM.2017.7992557

[10] Aleksander Skała, Zbigniew Waradzyn. Investigation and Determination of Efficiency of the Waste Heat Recovery System Using Peltier Modules. Conference: 2018 Conference on Electrotechnology: Processes, Models, Control and Computer Science (EPMCCS). DOI: 10.1109/EPMCCS.2018.8596493

Model of Hybrid Timetables for High Speed Urban Tramway Movement

Aleksei Gorbachev
Automatics and Remote Control on Railways
Emperor Alexander I St. Petersburg State Transport University
St. Petersburg, Russia
ag@agpage.ru

***Abstract*—Research review of urban traffic planning (including routes planning, distribution of vehicles between routes, timetabling and drivers assignment) using periodic and aperiodic timetables is given. Peculiarities of timetabling process in Russia and former USSR countries are described using the Public Transport Network and Event Transport Network presentation. The main goal of the research is creating of math model for solving the main task of schedules theory for high speed urban tramways with peculiarities of timetables mentioned of former Soviet countries. Tasks of this research are analyzing of previous works in this field, formalizing of subject area and creating of a math model which can be practically used, program implementation of the suggested model and analyzing its complexity. The concept of hybrid timetables is proposed. These timetables are similar to cyclic timetables inside the middle of one period of a day and they are similar to non-cyclic timetables inside period borders. Their usage helps to avoid most of limitations of non-cyclic schedules and it is convenient to use them for high-speed urban transport are defined in the article as a new class of schedules. Math model of hybrid timetables was suggested by the author. Polynomial complexity of the model was shown because it is analog of non-cyclic schedules. Software implementation of hybrid timetables model is described. Practical application recommendations of hybrid timetables usage are given on example of Automated System "Raspisanie Transporta" (AS RT), using example of the new high-speed tramway in St. Petersburg. Main tasks solved by this information system are described. Module structure of AS RT includes car distribution and trace planning modules, input-output module, core calculation module and graphical user interface module. Examples of hybrid timetables before and after departure intervals aligning are given to show the practical usage of the given math model. Interval diagrams by hours published for these examples.**

Keywords—timetable, schedule, urban transport, tramway, urban transport timetable, periodic timetable, cyclic timetable, non-cyclic timetable, aperiodic timetable, hybrid timetable

I. Introduction

Increasing of passenger traffic needs fast and qualitative urban transport movement planning. There are a lot of different mathematical models, technologies to organize work of public transport and different planning tools, created especially for these technologies. For example, there are models, invented for planning of railway transport work [1-4], air traffic [5-6], models for urban underground [7-9] or ground [10-12] transport and so on. But there is a few number of works devoted to work of public transport in case of movement in common traffic with all cars when run times depends on periods of a day [12, 15]. More detail overview of existing works in the field of timetables will be given below during the description of planning stages and classification of timetables from periodicity point of view.

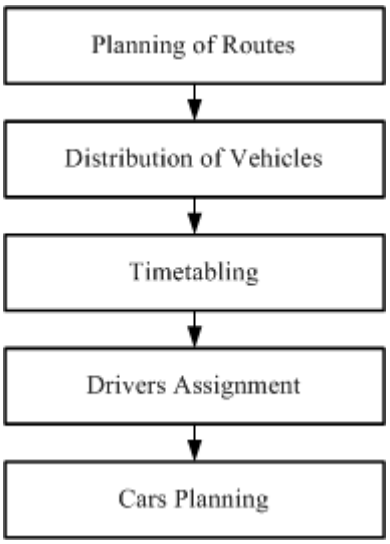


Fig. 1. Stages of planning process

Most of technologies and planning tools contain several successive stages, which will be described further in Fig 1.

II. Movement Planning Stages

A. Routes Planning

Planning of routes and lines means planning of standard traces.

Route in a path in a Public Transport Network (PTN) multi-graph. $PTN = (CP; P)$ is a multi-graph, which consists of finite number of control points $cp_i \in \{CP\}$ and finite number of segments, connecting these control points $p \in \{P\}$, where $p_i \in \{cp_i, cp_{i+1}\}$ (in Fig 2). Every segment connects neighboring control points. So control points are vertexes of this multi-graph and connecting segments are axes [14].

Usually route works between two end control points which are named end stations. Therefore, $PTN = (CP; P)$ can be presented as $PTN = (EndSt; P)$.

Only PTN for routes with the one common end station is covered by this article.

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

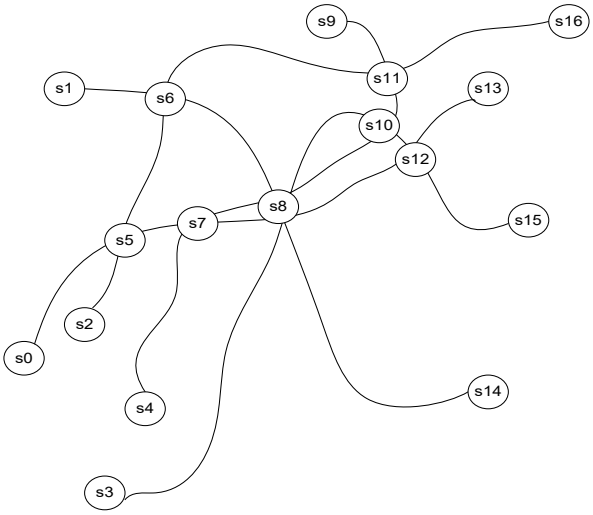


Fig. 2. Example of Public Transport Network

B. Distribution of Vehicles Between Routes

Distribution of rolling stock between routes (or frequency planning) is important part of planning. Distribution of rolling stock means calculating quantity of vehicles for each route. The frequency of a line shows number of trains are worked in a certain time period. Frequencies are important for passengers because they influence on waiting times in complex routes when passenger changes the route and waits for a vehicle on the new route. It is especially important in case of minimizing of total passengers' transfer time as a criterion for optimizing timetables [10,14].

But in this article another criterion of efficiency will be used. Therefore distribution of vehicles is interested only from timetabling aspect.

C. Timetabling

Next stage is timetabling. Timetabling is an important stage of movement planning. Timetable itself is the main document for every transport company because it declares number of races to be paid from a city budget. So this is the main document for transport company and also it is the main document for passengers who plan transfers from one place to another. This explains relevance of timetabling.

Timetabling contains two steps: generation of timetable's structure (races, setting different events such as driver changes and so on) and solving the main task of the schedule's theory: to find a feasible timetable [15]. Timetable is called feasible in general if it respects all capacity restrictions, lower and upper bounds for driving, changing and other stop limit times. So the main task of schedule's theory is setting arrival and departure times to the timetable of the known structure. Only the last problem (solving main task of the schedule's theory) is covered by this article.

Timetables' source data formalization is given taking into account Russian conditions is given in [15]. It is convenient to change presentation of data from PTN multigraph to an Event Activity Network (EAN). Events in case of route transport timetables are arrival times and departure times. EAN is graph of times, where each vertex means independent time value and each edge means a logical connection between time values (in Fig. 3).

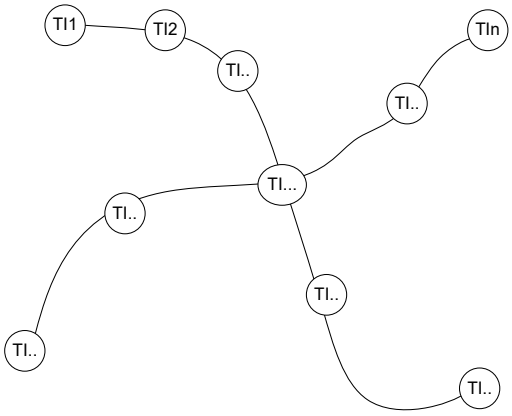


Fig. 3. Example of Event Activity Network

Last stages of urban transport planning process are setting of drivers to warrants and cars assignment between warrants. These stages are important for the transport company but less important to passengers. That's why this data is usually part only of internal business processes of the transport company.

D. Drivers assignment

Drivers assignment is the planning of real drivers to work on warrants of the current timetable. Driver assignment consists of two steps: dividing of warrants into several driver change parts (in general) and assigning of real drivers into driver change parts. Driver change part is part of driver change when one driver continuously works on one vehicle. If conductors also work on this route, a conductor change part is equal to those driver change part, who works with him.

Dividing of warrants into several driver change parts can be done during timetabling. It depends on operational technology which is used in the current transport company or even from the labor law of the current country and local regulations. For example, in Russia and former USSR countries it is popular to do this during timetabling. It is possible because of stable labor law, almost equal to all transport companies. This fact allowed to create several driver shift schedules, which are used on practice [15]. That's why it is easy to divide warrants to changes unlike situation in western countries, for example, in Germany, where dividing warrants into driver change parts depends on many local regulations of the current company [10]. In the last case it is very complicated and expensive to formalize this technology.

Second step is assigning of real drivers into change parts. The main idea of this process is distribution of change parts between drivers to follow the labor law and to make the total work time during the whole month nearer to normative value.

E. Cars Planning

Cars planning is the process of free cars setting into warrants. It depends on free rolling stock which is allowed to operate on the current line considering maintenance breaks.

III. Periodic Timetables

Periodic timetable (cyclic) are those timetables which have fixed period *Period* after which timetable repeats and which have limitations *[U,V]* which limits lower and upper values. These timetables appeared later then aperiodic timetables and they are supposed to be better for passengers because they are more predictable and it is easier to plan travels for passengers especially in case of complex trips with changing lines.

Math models for periodic timetables first was developed in [16] as Periodic Event Schedule Problem (PESP). PESP computational complexity is NP. It was proved in [16]. PESP can be solved by different ways which are better or worse for conditions of urban [7], railway [13] or air [5] transport.

Periodic timetables became very popular after increasing of computational powers and they are used worldwide, especially in Europe when transport companies want to optimize waiting times [7].

But some peculiarities of transport work in Russia (such as different run times during periods of a day, fixing of vehicles for drivers and so on) made using such timetables inexpedient or even impossible [15].

IV. Aperiodic Timetables

Aperiodic timetables are timetables which have no constant periods of repeat. Arrival and departure times can be calculated as independent values. Usually we create vector of independent values depending on limitations and restrictions. There is less restrictions on time limitations [U, V] comparing with periodic timetables.

Aperiodic timetables have polynomial complexity. It was first shown in [17] using flow approach. Polynomial complexity [17] is one of important benefits of aperiodic timetables from a practical point of their designing. Aperiodic timetables were popular in GDR, USSR and other countries during XX century. But changing of intervals and stops durations during a day, different events during a change (such as technical stops) make coordination of different routes difficult and unobvious for passengers.

V. Main Peculiarities of Timetabling Technology in Russia and Other Former USSR Countries

There are some peculiarities of timetabling technology in Russia and the other former USSR countries.

A. Different Run Times for Day Periods

The first feature is using different vehicle run times during period of day due to traffic jams in big cities. It is so because public transport work is usually organized in the same traffic flow where cars are moving. The following day periods are usually used: time before morning peak hours (from border of day till 7 a.m.), morning peak hours (from 7 a.m. till 9 a.m.), time before morning and evening peak hours (from 9 a.m till 4 p.m.), evening peak hours (from 4 p.m. till 8 p.m.), time after evening peak hours (from 8 p.m till end of a day). The exact borders are given only as an example, in practice they differ from one route to another.

B. Fixing Of Drivers to Vehicles During the Whole Change

The second feature is the fact that one driver and conductor usually work only on one vehicle during the whole change. It means that vehicles stand idle during driver and conductor dinners and other technological stays.

C. Logical Binding of Drivers to Vehicles

The third feature of these timetables is fixed binding of driver and conductors to one vehicle during long period (may be, for years). This specialty of former USSR countries is due to a large number of old different types of vehicles which are used in one transport company and each of them has their own traits and problems. This feature is usually called logical binding of transport teams on routes and vehicles. Each team consists of one driver and one conductor.

D. Optimization criterion

The fourth feature is an optimization criterion, which is used to rate the quality of timetables. The most typical one in the world is minimizing total passengers' travel time. It is a good criterion, but we need to know correspondence matrices and run times according to route traces to calculate total travel time. There is no any special equipment on vehicles to calculate passengers. So it is impossible to learn quickly changes in correspondence matrices.

The main practically used criterion of timetable optimization in such cases is uniformity of intervals. Usually departure intervals are used in this case simply because they are more important for passengers. The main hypothesis of this criterion choosing is that in general uniform distribution of departures (during one period of day) is more preferable for passengers than the others.

Aperiodic timetables are the most popular in Russia because they allow to reduce bad effects of some peculiarities which were mentioned above in section V. For example, flexible intervals' and stops' durations can avoid big stable intervals which appear in case of using periodic timetables. But it is less convenient for passengers, especially in case of composite trips.

The main goal of the research is creating of a math model for solving the main task of schedules theory for high speed urban tramways with peculiarities of former Soviet countries.

The first task is analyzing of previous works in this field was formalizing of timetabling process and previous experience which was described in sections I-IV.

The second task is creating of a math model which can be practically used for high speed urban transport with peculiarities mentioned in the section V.

The third task is program implementation of the suggested model and checking it on practice.

VI. Hybrid Timetables

A. New Routes

It is possible to use small stable intervals in case of technology with interchangeable drivers. Such planning technology can be applied for routes where all drivers are allowed to work on every vehicle during a day where this driver was set after distribution of vehicles. In this situation one change consists of several change parts. During every change part the driver works only on one vehicle continuously. This is usually possible only on routes with vehicles of one type which have no individual traits or problems.

B. Definition of Hybrid Timetables

Hybrid timetables which have stable intervals during one period of day are suggested in this article to avoid disadvantages of periodic timetables. For this types of timetable main intervals are fixed for each period during stage of vehicles' distribution between routes. Let us assume that main interval for period is fixed time between two nearest departure times from the same control point. Interval is constant during one period and changes only near period border to provide smooth transition to the main interval of the neighboring period.

Unfortunately, in most of cases it is impossible to avoid crossings with car traffic at the same level for ground urban transport in existing cities. That's why hybrid timetables use different run times for each period of a day.

C. Calculation of Main Intervals for Hybrid Timetables

Main intervals depends on trace run times, stop durations and number of vehicles used in each period of a day. We calculate number of vehicles working on current timetable on stage of vehicles' distribution between routes.

TABLE I. MAIN INTERVALS EXAMPLE

Route	Run Time, minutes	End. St.1 Stop, minutes	End. St.2 Stop, minutes	Total Race, minutes	Interval, minutes	Number of Vehicles
From start of work till 8 a.m.						
8	15,5	2	12	45	9	5
59	21,5	2	9	54	18	3
63	19,5	2	13	54	18	3
64	26,5	2	8	63	9	7

Let us consider a PTN with one common end station and a common part of the main trace of each route. For such type of timetables it is possible to use the following table to distribute vehicles between routes and to provide fixed interval for the common part.

D. Aligning of Intervals near Period Borders

We need to change interval near the each period border. It is possible to do discretely (for example to send fixed number of vehicles to park) or to do it smoothly step by step.

Aligning of intervals near period borders can be done using the same mathematical models that are used for aperiodic timetables.

In this model it was decided to do it smoothly by providing intervals aligning. So every part of such periodic timetable we analyze as as an individual aperiodic timetable between bordered with two ends *[Start, End]*.

To calculate vector of independent times *[TI]* we can generate standard grid. Standard grid is a vector with the same size *k* as *[TI],* in which time values are distributed inside $[TI_0; TI_k]$the most smooth way. So every value in *[Tin]* can be calculated according to the formula:

$$TIn_i = Round\left(TI_0 + TI_k - \frac{TI_0}{k \cdot i}\right) \quad (1),$$

where Round – is math round function to the nearest integer (in case of using only discrete time values in the timetable).

Vector *[Tin]* will be an ideal timetable from the aligning point of view, but excluding influence of *{U}* и *{V}*.

Optimized values of *[TI*] can be calculated using linear combination of *[TI]* и *[Tin]*. *s* – coefficient in this linear combination can be calculated the following way:

$$s = Min\left(s, \begin{cases} \dfrac{V_i - TIn_i}{TI_i - TIn_i}, TI > TIn_i \\ \dfrac{U_i - TIn_i}{TI_i - TIn_i}, TI < TIn_i \end{cases}\right) \quad (2)$$

After calculating s values of TI can be updated as a linear combination of old vector TI and ideal vector Tin:

$$[TI] = s[TI] + (1-s)[TIn] \quad (3)$$

Using liner programming for aligning of intervals is a typical thing for the aperiodic timetable model. This feature provides the polynomial complexity, which was shown in [10].

The best practical value of hybrid timetables is for routes with new rolling stock of one type where changing of vehicles is acceptable by drivers.

Model of hybrid timetables has polynomial complexity simply because it uses aperiodic model of aligning intervals.

But it has some limitations to use it in practice: intervals during different periods must be as close as possible to avoid long smoothing of intervals, transport company need to use technology with interchangeable drivers.

VII. PRACTICAL IMPLEMENTATION OF HYBRID TIMETABLES

A. Automated System "Raspisanie Transporta"

Creating of hybrid timetables was implemented in such planning tool as Automated System "Raspisanie Transporta" (AS RT). AS RT is a computer-aided design (CAD) system for building schedules in a graphical or table form for urban transport.. AS RT can be used independently or as a subsystem of City Electrical Transport Enterprise Resource Planning (ERP) System developed by IMSAT LTD Company which covers all stages of planning process shown in Fig. 1.

AS RT has a module structure and consists of several modules: car distribution module, trace planning module, core module, input-output (IO) library and graphical user interface (GUI) and form export module.

The main task of trace planning module is processing PTN data and supply them as initial data for scheduling.

The IO library provides serialization of timetable data to files or streams to read from hard disk or from database and write them to same destinations.

The core module works as a background during creating of timetables. It provides objects for different operations with timetables including elementary operations, schedule generation, intervals aligning, checks and so on.

Modules mentioned above work successively during timetable designing.

The GUI module works together with core library as a front-end. It is implemented via windows forms library and can be used both in Windows and Linux operation systems as a desktop application. This module provides two forms of representation of timetables: graphical picture and tables (see in Fig. 4).

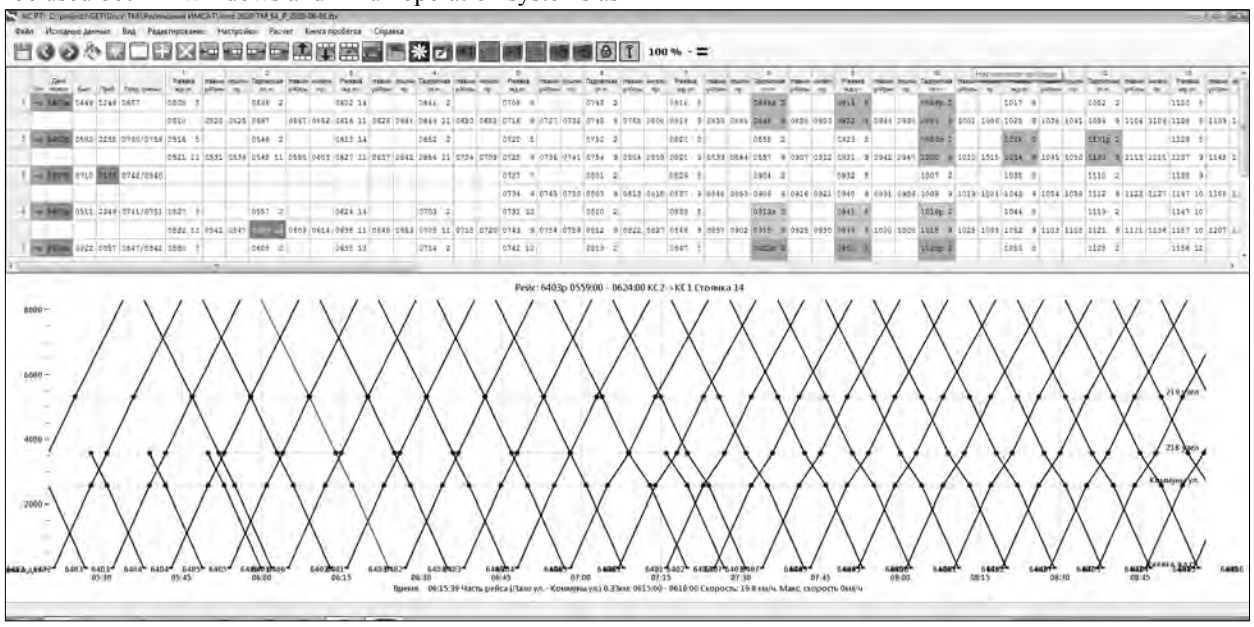


Fig. 4. "AS RT" main window interface

Form export module provides export of different documents (route timetable, driver timetable, control signs) to MS Excel format to print them or to send via e-mail.

AS RT supports three regimes for creating timetables.

1. Manual designing to create different types of timetables using operations which provides elementary operations. Different types of timetables can be created using these functions.
2. Automated operations – intervals aligning, recalculation of departures and arrivals for each stop for new run times. Different math models are used to provide aligning of various timetable types and other high-level automation functions.
3. Automatic designing – full generation of the whole timetable structure of races and times. This function provides fully automatic generation of some limited types of timetables which are the most popular on practice.

Intervals aligning in terms of AS RT means solving of main schedule's theory task (with calculating of limitations *{U, V}*). Suggested math model is a part of intervals aligning for hybrid timetables.

Before timetabling user need to run the special module for calculating of main intervals and distribution of vehicles between routes. User need to set the number of periods and their borders for each timetable, run time for each period of a day for the main trace, stop durations for end stations as initial data for work of this module. Than user can regulate main intervals for each period of a day to calculate the required number of vehicles. Results of calculations for four new routes in St. Petersburg for one period of a day are presented in table 1.

After calculation of main intervals user can start designing of the timetable for the each route.

Warrants can be added manually or initial version of timetable can be generated automatically.

B. *Practical Implementation for Planning Timetables on New Tramway Routes in St. Petersburg*

Let us check suggested formulas in section VI on example of existing timetable.

There is a part of route 59 timetable in table II. Warrant (column named "War.") in tables II and III means part of a timetable when one car works. Column named "Arr" means arrival time from a depot where every car starts it work.

Next columns named "St1" and "St2" mean end stations for a timetable. The first value of time in an each cell in a row for columns named "St1" and "St2" means an arrival time when car arrives at the end station 1 ("St1") or the end station 2 ("St2"). The second value of time in an each cell in a row for columns named "St1" and "St2" means a departure time when car departures from the end station 1 ("St1") or the end station 2 ("St2").

Departure intervals on end station1 (columns with name "St1") are presented in Fig. 5.

After aligning we will get Table III which is visualized in intervals diagram in Fig. 6. It will show that intervals only at 7 a.m. was changed. It means that an initial approximation was done quit well.

Generating initial approximation and timetable structure is beyond this article because it is not covered by the main task of schedule's theory.

Hybrid timetables of this type is possible to use when run times differ from period to period, but intervals are stable during one period.

Small fixed intervals are reasonable in planning technology with interchangeable drivers. Technology of work with interchangeable drivers means that drivers are not fixed on routes and vehicles during one change.

TABLE II. TIMETABLE PART BEFORE ALIGNING

War.	Arr.	St1	St2	St1	St2	St1	St2	St1	St2
5901L	0521	0521 0521	0542 0544	0605 0629	0650 0650	0712 0727	0749 0751	0813 0821	0843 0845
5902L	0521	0543 0543	0604 0606	0627 0650	0712 0712	0734 0745	0807 0809	0831 0839	0901 0903
5903L	0605	0605 0605	0626 0628	0649 0710	0733 0733	0755 0803	0825 0827	0849 0857	0919 0921

This is possible if driver is allowed to work on all types of vehicles on one route and vehicles do not have individual traits and problems. It is so on new routes where only new rolling stock is used.

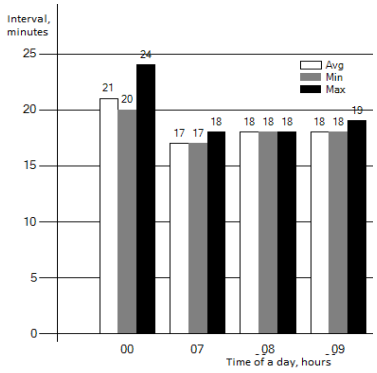


Fig. 5. Intervals diagram for Table II.

AS RT was implemented in St. Petersburg in new tramway company which was created by city government in partnership with the commercial company "LSR" to operate four new routes.

TABLE III. TIMETABLE PART AFTER ALIGNING

War.	Arr.	St1	St2	St1	St2	St1	St2	St1	St2
5901L	0521	0521 0521	0542 0544	0605 0629	0648 0650	0712 0727	0749 0751	0813 0821	0843 0845
5902L	0521	0543 0543	0604 0606	0627 0650	0710 0712	0734 0745	0807 0809	0831 0839	0901 0903
5903L	0605	0605 0605	0626 0628	0649 0710	0731 0733	0755 0803	0825 0827	0849 0857	0919 0921

It was built new independent infrastructure for these four routes and it bought new rolling stock of one type.

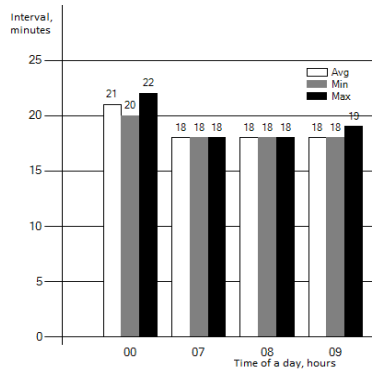


Fig. 6. Intervals diagram for Table III.

Using of new rolling stock made it possible to ensure the replacement of vehicles for one driver during a change.

AS RT provides manual and automated designing of the timetable's initial version. Hybrid timetables helped to organize synchronization between routes with almost fixed waiting times.

VIII. CONCLUSION

The scientific novelty of the article lies in the proposal to use hybrid timetables to decrease limitations of the aperiodic timetable model, especially for new high-speed urban public transport. Suggested model of hybrid timetables can be used to avoid problems with operating with almost stable intervals for different periods of a day.

The given model has polynomial complexity, so it can be used in practice without serious modifications of algorithms.

Inventing faster algorithms for such timetables is an important trend of further work in this field. But already implemented software demonstrates practical application of the suggested model.

Hybrid timetables have serious limitations comparing with existing aperiodic timetable model in the former USSR countries.

- Intervals during periods needs to be close to each other.
- Technology with fixing drivers is not acceptable for such timetables.
- Hybrid timetables cannot provide fully fixed intervals for passengers.

Despite given limitations hybrid timetables can provide quit easy model of different route synchronization for lines with one common part of a trace.

REFERENCES

[1] V. Cacchiani., A. Caprara, P. Toth, "Non-cyclic train timetabling and comparability graphs," Operations Research Letters, Elsevier, 38(3), 2010, pp. 179–184.

[2] F. Vautard, "Improvement of departure time suitability for interregional rail timetables", PhD thesis, KTH Royal Institue of Technology, Sweden, Stockholm, 2020, 37 p.

[3] C-W. Palmqvist, "Delays and Timetabling for Passenger Trains", PhD thesis, Lund University, Sweden, Lund: 2019, 107 p.

[4] S. Herrigel-Wiedersheim, "Algorithmic Decision support for the construction of periodic railway timetables", ETH Zurich, Zurich, 2015, pp. 100-120.

[5] X. Geng., M. Hu, "Simulated Annealing Method-Based Flight Schedule Optimization in Multiairport Systems", Mathematical Problems in Engineering Volume 2020, Article ID 4731918, https://doi.org/10.1155/2020/4731918.

[6] L. Lei, D. Zhao, H. Liu, D. Guo, "Flight Schedule Strategy of Airport Group". IOP Conf. Series: Materials Science and Engineering 790 (2020), UK, 2020, doi:10.1088/1757-899X/790/1/012102.

[7] C. Liebchen, "Periodic Timetable Optimization in Public Transport", PhD thesis, Technische Universitдt Berlin, Belrin, 2006, pp.20-25.

[8] C. Liebchen, M. Proksch, and F.H. Wagner, "Performance of algorithms for periodic timetable optimization", Computer-aided Systems in Public Transport, volume 600 of Lecture Notes in Economics and Mathematical Systems, Springer. Berlin Heidelberg, 2008, pp. 151–180.

[9] B. Illes, R. Ladanyi, G. Sarkozi, "Periodic timetable optimization in the public road transport services", University of

Miskolc, Advances Logistics Systems. Vol. 3(1), 2009, p. 219-225.

[10] M.E. Schmidt, “Integrating Routing Decisions in Public Transport Problems”, Springer Optimization and Its Applicaions 89, Springer Science + Business Media, New York, 2014, 386 p.

[11] E. Kochegurova, E. Gorokhova, “Optimizing Urban Public Transportation with Ant Colony Algorithm”, 8th International Conference on Computational Collective Intelligence, Greece: Halkidiki, 2016, - DOI: 10.1007/978-3-319-45243-2_45.

[12] A. Gorbachev, “Review of Urban Transport Timetbales Math Models”, Proceedings of Petersburg State Transport University, № 3. 15, PSTU, St. Petersburg, 2018, pp. 366-370.

[13] T. Lindner “Train Schedule Optimization in Public Rail Transport”, PhD thesis, Technische Universität Braunschweig, Braunschweig, 2000.

[14] Schöbel A., Schmidt M., “The Complexity of Integrating Routing Decisions in Public Transportation Models” –10th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS ’10), Liverpool, 2010, pp. 156-169.

[15] A. Gorbachev, “Urban Electrical Transport Timetable Synthesis Automation”, Proceedings of Petersburg State Transport University, № 4(41), PSTU, St. Petersburg, 2014, pp. 27-32.

[16] P. Serafani, W. Ukovich, “A mathematical Model for Periodic Scheduling Problems”, SIAM J. Disc. Math 2 (4), Society for Industrial and Applied Mathematics, USA, 1989, pp. 550-581.

[17] R.T. Rockafellar “Network Flows and Monotropic Optimization”, Athena Scientific, Belmont, 1998 – 634 p.

Using Additive Robust Modeling and Fault Simulation for Laser Ranging Measurements

Alexey Andreev
Institute of Physics
Kazan Federal University
Kazan, Russia
alexey-andreev93@mail.ru

Yury Nefedyev
Institute of Physics
Kazan Federal University
Kazan, Russia
star1955@yandex.ru

***Abstract*—This paper is focusing on the application of the regression modeling for analyzing highly accurate observations. In the study, the lunar laser ranging (LLR) measurements were assessed by robust analysis. Nowadays, statistical methods are applied in many branches of computer and information technology (astronomy, geology, ecology, geophysics, etc.). The use of the regression modeling (RM) allows producing not only reliable assessments for the parameters desired but also performing works on predicting the behavior of a system under consideration. Such investigations are conducted using both calculation algorithms of regression modeling (ARM) and the method of least squares (MLS) for assessing the quality of models. It is worth noting that the single use of the classic regression modeling method has certain disadvantages. The main problem is that the assumptions influence on the modeling results is not tested. As a result, the regression model may not correspond to the observations. The classic solution for this problem is checking the regression modeling conditions in case they are violated, one should apply other mathematic algorithms. This work suggests solving the problem described above by the additive dynamic regression modeling (ADRM). We developed the special software for ADRM which allows automatic checking the conditions reliability for RM-MLS systems and also conducting adaptation procedures if the conditions are not observed. To assess the model of observations, we use the method of simultaneous accounting of the presence of multicollinearity and violations of the normal error distribution. As a result, we produced assessments for the desired parameters with the minimum eliminating of erroneous observations. It should be noted that a significant performance loss is possible even at the limited eliminating of erroneous observations. To investigate the efficiency of LLR measurements, we produced the parameters as follows: percentage of erroneous observations eliminated, the standard value of the adaptation error, multi-parameter correlation, robust estimations of the ridge, its regression and stability.**

***Index Terms*—additive robust methods, regression analysis, software, multiple analysis**

I. Introduction

One of the most important computational procedures during observations processing is a stage of determining (estimating) parameters of the models used in geodesy, astrometry, and celestial mechanics. When describing processes or phenomena along with the problem of choosing a formal (approximating) or geometric (cause-and-effect) model there is also an important issue of determining correctness of mathematical processing, when the sample of the accepted data for processing and the used methods of applied mathematical statistics will not contradict the requirements for accuracy and reliability of the data obtained. Unfortunately, the conventional approach to estimating parameters during astronomical observations reduction, at which a rigidly fixed model and method of least squares (MLS) are used, does not correspond to the modern requirements of practice and the methodology capacity based on computer regression simulations. In the scientific practice there were attempts to go beyond the standard method of least squares [1], but they were focusing on solving particular problems and did not stipulate any systematic approach to solving the task. While processing space and ground-based observations, the typical limitations of using the MLS may be: 1) the presence of insignificant and doubling (dependent on each other) terms of the expansion; 2) violation of the MLS assumptions (normal Gauss-Markov scheme). The consequence of all this is an appearance of noise effects of various kinds that block up the description of the model and lead to as follows: 1) decrease in the accuracy of determining significant parameters; 2) do not allow reliable forecasting; 3) cause violation of the MLS main properties (consistency, unbiasedness, efficiency). The situation is also aggravated by the fact that during estimating parameters of the model its adequacy to observations is not controlled. Here, it is necessary to take into account that the applied set of quality measures is very narrow and the measures themselves have noticeable weaknesses. Since the observance of the MLS assumptions is not controlled within the classic approach, researchers most often do not know about the actual state of affairs and avoid using adaptive computational schemes. As an alternative to the classic approach, the present work suggests the methodology of regression simulation which uses regression analysis to solve the tasks of estimating the parameters of astronomical processes models. The regression analysis method allows: 1) testing the assumptions during construction and study of the model; 2) adaptation if the MLS conditions are not observed; 3) special software packages which are systems of processing information allowing to automate the processes of calculation and results analysis. The regression simulation is a systematic approach to analyzing models of astronomical processes which allows using any element of analysis of the system studied (sample, model, method of estimating parameters, method of estimating structures, quality measure, set of assumptions) cor-

rectly. At the same time, the chosen element of the regression model may be checked for reliability and if the specified conditions are not observed, the corresponding adaptation will be applied.

Time series (TS) analysis methods are a successive structural and parametric identification of TS multi-component models, providing an estimation of the built models quality on the accuracy of approximation and the prediction, and diagnostics of the conditions of the ordinary least-squares method (LSM). The basic method of its application involves the following steps: 1) graphic representation and description of the TS dynamics; 2) studying the TS properties by means of autocorrelation, spectral and wavelet analysis; 3) selection and removal of the trend and polyharmonic components; 4) investigating TS components, which remained after the identification of the above components; 5) building of a mathematical model to describe the random component (autoregressive model, moving average, martingale approximation, etc.), verification of its adequacy; 6) diagnostics of the LSM basic conditions; 7) analysis of the models quality; 8) forecasting the process development, represented by TS; 9) joint processing of TS (correlation and cross-correlation analysis). In this paper, additive modeling methods are used to solve this problem.

II. Subject and Method of Research: Additive Modeling and Computing Algorithms

Additive modeling was suggested by Jerome H. Friedman and Werner Stuetzle (1981) [2] and refers to non-parametric regression methods [3]. Additive modeling is based on the property of quantities, which is that an entire object can be represented as a certain sum of individual quantities, if this object allows it to be broken into its component parts [4]. For example, the additivity of the observed parameter means that this parameter is equal to the sum of the constituent parts of the given parameter. In other words, additive models can be used in cases where the final value is the algebraic sum of a series of factor values. Additive modeling has a greater ability to adapt than linear modeling, since it adapts better to the approximation errors than a general regression hypersurface [5]. The regression hypersurface is a figurative representation of regression equations. Here it should be noted that adaptive modeling does not take into account the emergent properties of the system and its qualitative basis [6]. Therefore, in this paper we have solved the problem by developing a method for improving the existing least squares estimation (LSE) algorithms so that the final results of calculations are less dependent on emissions.

The algorithm is a hybrid method based on Huber's method algorithms [6] and ridge regression, as described in [8].

Let us consider a linear model:

$$Y = X\beta + \varepsilon. \tag{1}$$

For system (1), elements of the vector of parameters are estimated by the formula:

$$\beta(\lambda) = \left(X^T W X + \lambda D\right)^{-1} X^T W Y, \tag{2}$$

where ($\lambda > 0$) is calculated by one-dimensional optimization; D– diagonal matrix elements $d_{ii} = a_{ii}$, $i = \overline{1,m}$ (diagonal elements of $A = X^T X$ are taken as the diagonal elements of D); W – diagonal matrix consisting of weights w_t:

$$w_t = \psi[(y_t - x_t a/S)]/[(y_t - x_t a)/S], \tag{3}$$

where w_t – weight attached to observation t; selection function ρ and therefore $\psi(z)$ is carried out by the method of Huber [7].

To ensure stability of the parameters of outliers [6] instead of minimizing the sum of squared deviations proposed to minimize the sum of less rapidly increasing functions:

$$\sum \rho(\varepsilon_i) \to \min \tag{4}$$

the contribution of the values $|\varepsilon_i|$, smaller in absolute value c, a measured in squares of deviations (as in LSM), but if $|\varepsilon_i| > c$, the contribution is measured in proportion $|\varepsilon_i|$.

Or solve the system:

$$\sum \psi(\varepsilon_i) x_{ik} = 0, i = 1, \ldots, n; k = 1, \ldots, p-1, \tag{5}$$

where $\psi = \rho'$ is chosen in order to ensure minimal dispersion matrix of estimates. In [6], this function is offered as:

$$f(n) = \begin{cases} -c, & \text{for } \varepsilon_i < -c \\ \varepsilon_i & \text{for } |\varepsilon_i| \le c \\ c & \text{for } |\varepsilon_i| > c \end{cases} \tag{6}$$

III. Algorithm of the Interactive Automated System for Additive Dynamic Regression Modeling and its Application in Work

In the work, two statistical measures of compliance are implemented: external - for a model that pretending to determinism and internal for the approximate model. The software package library status of compliance of assumptions is a set of procedures for implementing the verification of compliance with the following basic assumptions of regression analysis: redundancy and underdetermined model, multicollinearity of covariates model, the normality of the distribution of the model residues expectation, heteroskedasticity, independence of residuals.

When developing the software, visual programming Turbo Delphi is used. Embedded script allows the data processing with up to 10 times reduced time compared to interactive one. A user does not need to independently move from one procedure to another and there is no need to load a file with a control sample several times. The resulting optimal model has no violations.

ADRM is used for the analysis of various parameters of planetary models: gravitational and magnetic fields, physical surface reliefs on both the entire sphere and its separate parts using measurements series in the points with known coordinates, lunar laser ranging. The application of this software package to the analysis of space and ground-based observations allows for the construction of regression models and determination of lunar laser ranging observed parameters. There is an opportunity to assess the reliability and observance

quality of calculational procedures for the method of least squares. If these conditions are violated, a special algorithm of adapting the obtained results is activated. This calculational complex has been used in the construction of regression models.

The script is implemented as a separate module Automatic-Scheme.pas which is embedded in the package ADRM. When one loads the data file, it reads the information on the number of covariates and the number of observations. One should also select a folder to save the file with the results. There are two modes of the script: with automatic formation of a control sample and a sample taken from a file. The control sample is 10% of the total number of observations. The file name is entered in all the procedures covered in this scenario. At each stage of passing the script file intermediate results are formed, and during the search for optimal model the intermediate results are compared according to the criterion and the optimal model for the entire sample is recalculated.

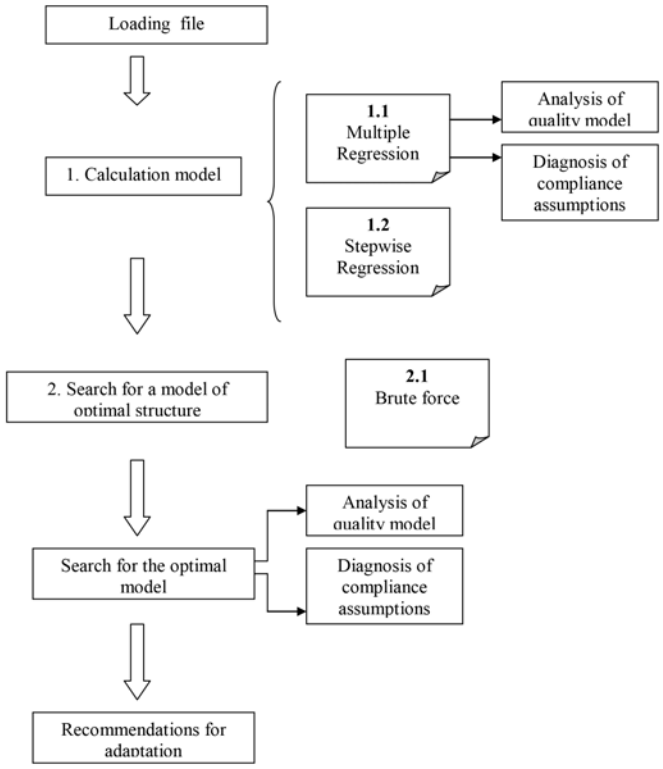


Fig. 1. Algorithm automatically follows the stages of data processing

If a user needs to analyze the model in terms of compliance with the assumptions of RA, as well as to refine the details of external and internal measures values, it is necessary to choose the analysis mode so that there are two stages modeling. As a result, we obtain the model with the multiple regression and optimal structure having a minimum estimate of random error. It was calculated by the formula:

$$\sigma_{\Delta} = \sqrt{\sum_{i=1}^{k} (\Delta_i - \overline{\Delta})^2 / (k - p)} \qquad (7)$$

where $\Delta_i = (y_i - \hat{y}_i)$ is the difference between the experimental and calculated values of the response; k is the number of control points; p is the number of covariates analyzed.

The received software package was tested on the data of Laser Ranging of the Moon (LRM). The following items are of interest: Lunar Moment of Inertia – Tracking data on orbiting spacecraft gives the 2-degree gravity harmonics J_2, C_{22}. From LLR one obtains the moment of inertia combinations $(C - A)/B$ and $(B - A)/C$. Combining the two sets gives C/MR^2, the polar moment normalized with the mass M and radius R. LRM is sensitive to the moment of the solid moon, without fluid core. Elastic Tides: Elastic tidal displacements are characterized by the lunar (2nd-degree) Love numbers – h_2, l_2. Tidal distortion of the 2nd-degree gravity potential and moment of inertia depend on k_2. Love numbers depend on the elastic properties of the interior including the deeper zones where the seismic information is weakest. LRM detects tidal displacements; determination of k_2 is potentially more accurate but is complicated. Tidal Dissipation: The tidal dissipation Q is a bulk property that depends on the radial distribution of the material Q_s. LRM detects four dissipation terms and infers a weak dependence of tidal Q on frequency. The tidal Q_s are surprisingly low, but LRM does not distinguish the location of the low-Q material. At seismic frequencies, low-Q material, suspected of being a partial melt, was found for the zone just above the core. Dissipation at a Liquid-Core/Solid-Mantle Interface: A fluid core does not share the rotation axis of the solid mantle. While the lunar equator precesses, a fluid core can only weakly mimic this motion. The resulting velocity difference at the core-mantle boundary causes a torque and dissipates energy. Several dissipation terms are considered in the LRM analysis in order to separate core and tidal dissipation. Applying Yoder's turbulent boundary layer theory yields upper limits for the fluid core radius. Inner Core: A solid inner core might exist inside the fluid core. Gravitational interactions between an inner core and the mantle could reveal its presence. Too little is known about the inner core to predict the size of the perturbation of the physical librations. Our work will aid these studies. Core Oblateness: Core oblateness influences solutions for the Love number k_2. Fluid Core Moment of Inertia is potentially detectable, but the present uncertainty is too large to be useful. A core radius of ~ 390 km is indicated if it is iron or larger if there are significant amounts of sulfur mixed in. A longer span of very accurate data is needed. Our efforts will advance LRM to mm-level range sensitivity and would allow for vastly improved accuracy of physics parameters. Anticipated improvements in Earth geophysics and geodesy results would include the positions and rates for the Earth stations, Earth rotation, precession rate, nutation, tidal influences on the orbit, and etc. The following lunar science questions can be discussions: What is the deep interior structure and properties? What are the core properties? Is there an inner core? What causes strong tidal dissipation? What roles did tidal and core dissipation play in the dynamical and thermal evolution? What stimulates free librations? This can be achieved with using additive modeling.

IV. RESULTS OF THE METHOD SOFTWARE ANALYSIS

The program was designed in a complex visual programming Turbo Delphi. The program's interface features a simple, clear presentation and easy navigation.

The numerical test the effectiveness of the methods of empirical data, as the main criterion of efficiency studied the standard deviation (SD) "Sigma Delta", obtained by the control sample and treated as external measure.

TABLE I
THE RESULTS ACCORDING TO "LUNAR LASER RANGING NUMBER OF OBSERVATIONS OF THE MOON" AT A RANDOM CONTROL SAMPLE

	%	σ	R	F	$\sigma\Delta$
RR		$1.1540 \cdot 10^{-8}$	0.873226	69.5708	$1.5468 \cdot 10^{-8}$
MH	5	$1.1694 \cdot 10^{-8}$	0.873227	66.9164	$1.5088 \cdot 10^{-8}$
SRR	5	$1.1467 \cdot 10^{-8}$	0.873226	69.7046	$1.4976 \cdot 10^{-8}$
MR		$1.1710 \cdot 10^{-8}$	0.873830	62.7108	$1.6297 \cdot 10^{-8}$

To test and analyze the effectiveness of the laser investigated a number of lightlocating observations of the Moon, previously treated in [6]. In tables 1,2 except "$\sigma\Delta$" are: % – the percentage of "soiling" sample, σ – the standard error of the approximation, R – multiple correlation coefficient, F – observed value of F-statistics, used the notation methods: RR – ridge regression, MH – Huber's method, SRR – stable ridge regression, MR – multiple regression (one of the computational schemes of LSM).

TABLE II
THE RESULTS ACCORDING TO "LUNAR LASER RANGING NUMBER OF OBSERVATIONS OF THE MOON" WITH A FIXED CONTROL SAMPLE (10% FROM THE END OF THE SERIES)

	%	σ	R	F	$\sigma\Delta$
RR		$1.0690 \cdot 10^{-8}$	0.873226	69.5708	$1.3280 \cdot 10^{-8}$
MH	5	$1.0805 \cdot 10^{-8}$	0.873227	66.9164	$0.9497 \cdot 10^{-8}$
SRR	5	$1.0573 \cdot 10^{-8}$	0.873226	69.7046	$0.9385 \cdot 10^{-8}$
MR		$1.1145 \cdot 10^{-8}$	0.882079	68.0712	$2.7608 \cdot 10^{-8}$

V. CONCLUSION

In this paper we have created a special automated software package for additive regression modeling that allows us to verify compliance with the assumptions between the parameters of the regression analysis and least squares estimation and to perform adaptation in the event of violations. The method of joint solution of least squares estimation for the case of violation of the linearity of error distribution and the absence of multicollinearity has been developed. This made it possible to improve the accuracy of estimating the model parameters and the forecast in the presence of emissions and multicollinearity using ridge regression estimations. Based on the goals, were solved three tasks: 1) The synthesis algorithm of the method of sustainable ridge estimation, adapted together to marked disturbances; 2) The software implementation of the new method; 3) The testing the program on empirical data. As result when analyzing the quality of the model for external as Sigma Delta method (Table 2) is stable ridge regression provides the highest accuracy of prediction.

It should be noted the initial assumptions of the regression analysis are always observed. However, discovering that the preconditions are violated is not sufficient. A specific software package containing particular measures that come into force under these conditions are required. Thus, for the effective use of adaptive regression modeling approach (ARM) one should the apply a particular software package to automate the process of taking observations, analyze the quality of the models produced and analyze the compliance with the assumptions of regression analysis using the ordinary least squares method (LSM), as well as implement the appropriate procedures to adapt. The purpose of this study is to improve the performance of the computational modeling process by automating the search for the optimal set of regressors, and analyze it. To achieve this goal, it is necessary to solve a number of problems: 1) Development of the software package "Interactive Automated System for Optimal Regressions Modeling" (IASORM) based on connecting library quality analysis model with the compliance status of assumptions; 2) Implementation of the algorithm scenario of automatic data processing with the functional connection of libraries. The software package IASORM is a specialized system that implements the strategy of regression modeling. Automated script processing can improve the effectiveness of the existing methods of the package. Embedded library of the quality analysis and of the compliance model assumptions extend functionality for the user and is aimed at identifying the adequacy of models and observations in order to detect violations of the basic assumptions of regression analysis. Proposed scenario increases the computational process speed compared to interactive computing. IASORS implements the strategy of statistical (regression) modeling. This software package can be used to create regression models and predict dynamic processes.

The results of the work will be used in the reduction of lunar observations [9], [10] and other positional observations [11].

ACKNOWLEDGMENT

This work was partially supported by Russian Science Foundation, grants no. 20-12-00105 (according to the grant, the method for data analysis was created) and 19-72-00033 (according to the grant, the numerical calculations were carried out). This work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University. This work was partially supported by a scholarship of the President of the Russian Federation to young scientists and post-graduate students SP-3225.2018.3, the Russian Foundation for Basic Research grant no. 19-32-90024 and the Foundation for the Advancement of Theoretical Physics and Mathematics "BASIS".

REFERENCES

[1] S. G. Valeev, Regression modeling at data processing. Kazan: FAN, 2001.

[2] J. H. Friedman and W. Stuetzle, “Projection pursuit regression,” J. Am. Stat. Assoc., 1981, vol. 76, pp.817–823.
[3] M. Y. Cheng and J. Fan, “Peter Halls contributions to nonparametric function estimation and modeling,” Ann. Stat., 2016, vol. 44, iss. 5, 1837–1853.
[4] M. Yuan, “On the identifiability of additive index models,” Stat. Sin., 2011, vol. 21, iss. 4, pp. 1901–1911.
[5] Y. A. Nefedyev, V. M. Bezmenov, S. A. Demin, A. O. Andreev and N. Y. Demina, “Application of antijamming robust analysis method for selenocentric reference net building,” Nonlin. Phen. in Complex Syst., 2016, vol. 19, iss. 1, pp. 102–106.
[6] Y. A. Nefedjev and N. G. Rizvanov, “The results of an accurate analysis of EAO charts of the Moon marginal zone constructed on the basis of lunar occultations,” Astron. Nachr., 2002, vol. 323, iss. 2, pp. 135–138.
[7] S. G. Valeev, “Regression modelling in selenodesy,” Earth Moon Planets, 1986, vol. 35, iss. 1, pp. 1–5.
[8] S. G. Valeev, “Coordinates of the Moon reverse side sector objects,” Earth Moon Planets, 1986, vol. 34, iss. 3, pp. 251–271.
[9] N. Rizvanov and J. Nefedjev, “Photographic observations of Solar System bodies at the Engelhardt astronomical observatory,” Astron. Astrophys., 2005, vol. 444, iss. 2, pp. 625–627.
[10] N. Y. Varaksina, Y. A. Nefedyev, K. O. Churkin, R. R. Zabbarova and S. A. Demin, “Selenocentric reference coordinates net in the dynamic system,” J. Phys. Conf. Ser., 2015, vol. 661, iss. 1, p. 012014.
[11] N. G. Rizvanov, Y. A. Nefed’ev and M. I. Kibardina, “Research on selenodesy and dynamics of the Moon in Kazan,” Sol. Syst. Res., 2007, vol. 41, iss. 2, pp. 140–149.

Integrated-Optics Quantum Processor Based on Entangled Photons in Coupled Cavities

Farid Ablayev
Kazan Federal University
Kazan, Russian Federation
fablayev@gmail.com

Alexander Vasiliev
Kazan Federal University
Kazan, Russian Federation
vav.kpfu@gmail.com

Sergey Andrianov
Tatarstan Republic Academy of Sciences, Institute for Applied Research
Kazan, Russian Federation
andrianovsn@mail.ru

Sergey Moiseev
Kazan Quantum Center of Kazan National Technical University named after A.N. Tupolev - KAI
Kazan, Russian Federation
samoi@yandex.ru

***Abstract* — We propose a scheme of the universal quantum processing unit based on integrated optic waveguide excitation transfer of qubits between optical micro-resonators and Kerr nonlinear interaction between neighboring cavities. We present the protocols for the implementation of single-qubit gates and a two-qubit controlled phase gate. The optimal regimes of gates operation are studied using input-output formalism.**

Keywords — quantum processor, qubit, quantum gate, memory, input-output formalism

I. Introduction

Quantum gates on cross-Kerr nonlinearity of photon qubits are well-known [1,2]. Essential cross-Kerr nonlinearity was obtained experimentally in the paper [3]. However, it was shown by Shapiro that cross-talk noises can be too high in the case of local nonlinearity [4]. These restrictions were theoretically overcome in papers [5,6] for physical photon qubits interacting with the set of distributed optical cavities coupled with atoms. Here, we will consider the Controlled Phase gate (CPHASE, for short) on logical photon qubits with the use of optical cavities coupled with atoms. Being essentially nonlocal, these qubits are able to allow the separation of signal from noises using Purcell effect in separate cavities.

The CPHASE gate is the main entangling gate of our set of operations, constituting universal basis for arbitrary quantum computations. Due to the pairwise logical encoding the proposed model may reduce the number of elementary gates for implementing complicated quantum algorithms and provides resistance to a number of quantum errors though at the price of doubling the number of physical qubits required.

II. Quantum Processor Architecture

Let us consider the array of photon micro-cavities with cross-Kerr nonlinearity between them constituting quantum processing unit (Fig.1).

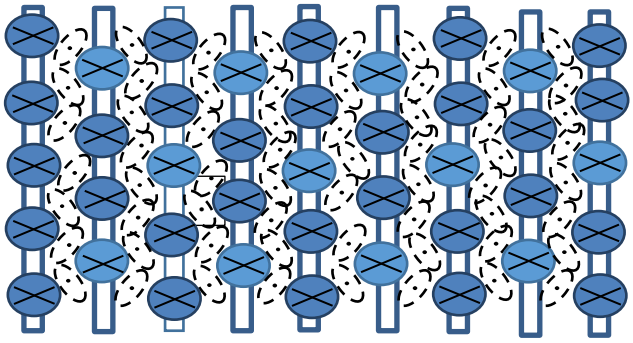


Fig. 1. Scheme of Quantum Processing Unit (QPU) on the arrays of three micro-cavities coupled via two-level atoms.

We use the pairwise logical encoding $|0_L\rangle = |01\rangle, |1_L\rangle = |10\rangle$ for operation of all gates. The CPHASE operation transforms target logical pair depending on the control logical pair. So for this operation we put a single photon excitation of the target pair into left and right cavities in a three-cavity array, while the middle cavity holds the photon excitation of one of the control qubits transferred from optical memory via nearby cavities coupling through common atoms between them. They can be moved to resonance or removed from it by an external field.

III. Controlled Phase Gate.

A. Basic operation mode.

Let us now consider the two-qubit CPHASE operation. First, we prepare initial state of the target qubit by loading excitation to the side cavities of selected three-qubit array (Fig.2) in the way described above. Then, we load excitation from one of the cavities of another three-qubit array to the middle control cavity of selected array and switch off further transfer of excitation among cavities. In addition, we switch off interaction between the first two cavities in the array.

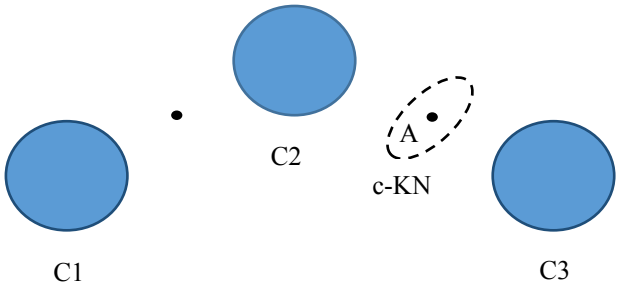


Fig. 2. Physical scheme of photon qubit CPHASE gate on the array of three micro-cavities coupled by cross-Kerr nonlinearity (C1, C2 and C3 are micro-cavities, A is an atom, c-KN is cross-Kerr nonlinearity).

Then, the Hamiltonian of the system can be written as

$$H = H_0 + H_1, \tag{1}$$

where

$$H_0 = \hbar\omega_1 n_1 + \hbar\omega_2 n_2 + \hbar\omega_3 n_3 \tag{2}$$

is main Hamiltonian, ω_i is frequency of photons in cavity i, n_i is number of photons in cavity i,

$$H_1 = \chi_{12} n_1 n_2 + \chi_{23} n_2 n_3 \tag{3}$$

is perturbation Hamiltonian, χ_{12} and χ_{23} is third order optical susceptibility. The wave function can be written as

$$\psi = c_1|1\rangle_1|0\rangle_2|0\rangle_3 + c_2|0\rangle_1|0\rangle_2|1\rangle_3 + c_3|1\rangle_1|1\rangle_2|0\rangle_3 + c_4|0\rangle_1|1\rangle_2|1\rangle_3, \tag{4}$$

where $|n\rangle_i$ is the state of n photons in cavity $i = 1,2,3$.

Writing down the Schrodinger equation $\frac{d\psi}{dt} = -\frac{i}{\hbar}H\psi$, we have

$$\frac{dc_1}{dt} = -i\omega_1 c_1, \tag{5}$$

$$\frac{dc_2}{dt} = -i\omega_3 c_2, \tag{6}$$

$$\frac{dc_3}{dt} = -i\left(\omega_1 + \omega_2 + \frac{1}{\hbar}\chi_{12}\right)c_3, \tag{7}$$

$$\frac{dc_4}{dt} = -i\left(\omega_2 + \omega_3 + \frac{1}{\hbar}\chi_{23}\right)c_4. \tag{8}$$

The solutions of these equations are the following:

$$c_1 = e^{-i\omega_1 t}, \tag{9}$$

$$c_2 = e^{-i\omega_3 t}, \tag{10}$$

$$c_3 = e^{-i\left(\omega_1+\omega_2+\frac{1}{\hbar}\chi_{12}\right)t}, \tag{11}$$

$$c_4 = e^{-i\left(\omega_2+\omega_3+\frac{1}{\hbar}\chi_{23}\right)t}. \tag{12}$$

If $\omega_1 t = 2\pi m_1$, $\omega_2 t = 2\pi m_2, \omega_3 t = 2\pi m_3$, $\chi_{12} = 0$ we have for transfer matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{-i\varphi} \end{pmatrix}, \tag{13}$$

where $\varphi = \frac{1}{\hbar}\chi_{23}t$. This transfer matrix corresponds to CPHASE gate.

However, the gate given by the matrix (14) is physical one, and we need to demonstrate its operation in the Hilbert subspace generated by the logical encoding $|0\rangle_L = |01\rangle, |1\rangle_L = |10\rangle$. Fortunately, this can be easily done by applying this two-qubit gate to the first qubit of each logical pair.

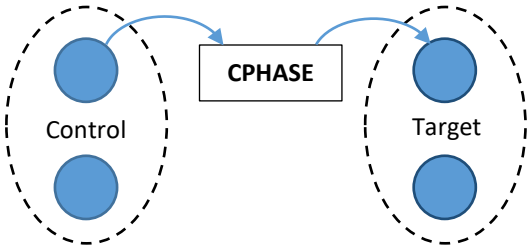


Fig. 3. Logical scheme of the CPHASE gate, that physically operates on the first qubit of each logical pair.

Consider the general case of the logical control and target pair:

$$|\psi_C\rangle|\psi_T\rangle = (a_1|0\rangle_L + b_1|1\rangle_L) \otimes (a_2|0\rangle_L + b_2|1\rangle_L) = a_1a_2|0101\rangle + a_1b_2|0110\rangle + b_1a_2|1001\rangle + b_1b_2|1010\rangle. \tag{14}$$

Since we are acting on the first and third qubit of the four-qubit physical state and the only nontrivial result of the CPHASE gate is for the basis state $|11\rangle$, we obtain the following result:

$$|\psi_C\rangle|\psi_T\rangle \to a_1a_2|0101\rangle + a_1b_2|0110\rangle + b_1a_2|1001\rangle + b_1b_2e^{-i\varphi}|1010\rangle = a_1a_2|0\rangle_L|0\rangle_L + a_1b_2|0\rangle_L|1\rangle_L +$$

$$b_1 a_2 |1\rangle_L |0\rangle_L + b_1 b_2 e^{-i\varphi} |1\rangle_L |1\rangle_L ,$$

(15)

which is exactly the desired action of the CPHASE gate in the logical subspace we use.

B. LOSSES IN INPUT-OUTPUT FORMALISM.

Let's take into account the coupling of cavities with surrounding medium using input-output formalism. In its framework, we have the following equations on photon annihilation operators

$$\frac{da_n}{dt} = -\frac{i}{\hbar}[a_n, H] - \frac{\gamma}{2} a_n + \sqrt{\gamma} a_{n,IN}, \quad (16)$$

where $n = 1,3$ is the number of cavity, γ is cavity surrounding medium coupling parameter. With that, we assume the second cavity not to be coupled with its feeding waveguide in moment t_0 since it was excited with quantum probability beforehand, previously to excitation of cavities 1 and 3. We have

$$\frac{da_1}{dt} = -i\omega_1 a_1 - \frac{\gamma}{2} a_1 + \sqrt{\gamma} a_{1IN}, \quad (17)$$

$$\frac{da_2}{dt} = -i\omega_2 a_2 - \frac{\gamma_c}{2} a_2, \quad (18)$$

$$\frac{da_3}{dt} = -i\omega_3 a_3 - \frac{i}{\hbar}\chi_{23} a_3 n_2 - \frac{\gamma}{2} a_3 + \sqrt{\gamma} a_{3IN}. \quad (19)$$

Values n_2 and n_3 in equations (18) and (19) are not constant in time. In order to evaluate the rate of this temporal variation, let us compose equation on n_3. It can be written as

$$\frac{d}{dt}(n_3) = -\gamma n_3 + \sqrt{\gamma}(a_{3IN}^+ a_3 + a_3^+ a_{3IN}). \quad (20)$$

Integration of equation (20) yields

$$n_3(t) = n_3(0) - \delta n_3(t), \quad (21)$$

where

$$\delta n_3(t) = \int_0^t \{\gamma n_3 - \sqrt{\gamma}(a_{3IN}^+ a_3 + a_3^+ a_{3IN})\}\, dt'. \quad (22)$$

If γ is small than $n_3(t) \approx n_3(0)$. Similarly, it can be shown that $n_2(t) \approx n_2(0)$.

After Fourier transform Equation (17) yields the following relation

$$a_1(\omega) = \frac{\sqrt{\gamma}}{i(\omega_1 - \omega) + \frac{\gamma}{2}} a_{in,1}(\omega). \quad (23)$$

Starting from input-output formalism relation

$$a_{out,1}(\omega) = \sqrt{\gamma} a_1(\omega) - a_{in,1}(\omega), \quad (24)$$

we get using (23) the following expression

$$a_{out,1}(\omega) = \frac{-i(\omega_1 - \omega) + \frac{\gamma}{2}}{i(\omega_1 - \omega) + \frac{\gamma}{2}} a_{in,1}(\omega), \quad (25)$$

Analogously, we have

$$a_{out,3}(\omega) = \frac{\frac{\gamma}{2} - i\left(\omega_3 + \frac{1}{\hbar}\chi_{23} n_2(0) - \omega\right)}{\frac{\gamma}{2} + i\left(\omega_3 + \frac{1}{\hbar}\chi_{23} n_2(0) - \omega\right)} a_{in,3}(\omega). \quad (27)$$

We can write for input wave function

$$\begin{aligned}\psi_{IN}(\omega) = {} & c_1 a_{in,1}^+(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} \\ & + c_2 a_{in,3}^+(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} \\ & + c_3 a_{in,1}^+ a_2^+(t_0)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + \\ & c_4 a_{in,3}^+(\omega) a_2^+(t_0)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3}. \end{aligned} \quad (28)$$

In the same way, we write for the output wave function

$$\begin{aligned}\psi_{OUT}(\omega) = {} & c_1 a_{out,1}^+(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} \\ & + c_2 a_{out,3}^+(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} \\ & + c_3 a_{out,1}^+(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} \\ & c_4 a_{out,3}^+(\omega) a_2^+(t_0)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3}. \end{aligned} \quad (29)$$

Using equations (25,27) we have

$$\begin{aligned}\psi_{OUT} = {} & c_1 \frac{\frac{\gamma}{2} + i(\omega_1 - \omega)}{\frac{\gamma}{2} - i(\omega_1 - \omega)} a_{in,1}^+(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + \\ & c_2 \frac{\frac{\gamma}{2} + i(\omega_3 - \omega)}{\frac{\gamma}{2} - i(\omega_3 - \omega)} a_{in,3}^+(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + c_3 \frac{\frac{\gamma}{2} + i(\omega_2 - \omega)}{\frac{\gamma}{2} - i(\omega_2 - \omega)} \cdot \\ & \frac{\frac{\gamma}{2} + i(\omega_1 - \omega)}{\frac{\gamma}{2} - i(\omega_1 - \omega)} a_{in,1}^+(\omega) a_2^+(t_0)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + \\ & c_4 \frac{\frac{\gamma}{2} + i\left(\omega_3 + \frac{1}{\hbar}\chi_{23} - \omega\right)}{\frac{\gamma}{2} - i\left(\omega_3 + \frac{1}{\hbar}\chi_{23} - \omega\right)} \cdot \\ & \frac{\frac{\gamma}{2} + i(\omega_2 - \omega)}{\frac{\gamma}{2} - i(\omega_2 - \omega)} a_{in,3}^+(\omega) a_2^+(t_0)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3}. \end{aligned} \quad (30)$$

If the spectrum bound width of photon is narrow, we can assume $\omega_1 = \omega_2 = \omega_3 = \omega$ and, consequently, we have

$\omega_3 + \frac{1}{\hbar}\chi_{23} - \omega \approx \frac{1}{\hbar}\chi_{23}$. In this case, equation (30) yields the following expression:

$$\psi_{OUT}(\omega) = c_1 a^+_{in,1}(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + c_2 a^+_{in,3}(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + c_3 a^+_{in,1}(\omega)a^+_2(t_0)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + c_4 \frac{\frac{\gamma}{2}+i\frac{1}{\hbar}\chi_{23}}{\frac{\gamma}{2}-i\frac{1}{\hbar}\chi_{23}} a^+_{in,3}(\omega)a^+_2(t_0)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3}. \quad (31)$$

Taking into account that

$$\frac{\frac{\gamma}{2}+i\frac{1}{\hbar}\chi_{23}}{\frac{\gamma}{2}-i\frac{1}{\hbar}\chi_{23}} = e^{2iarctg\{\frac{2}{\hbar\gamma}\chi_{23}\}}, \quad (32)$$

we have

$$\psi_{OUT}(\omega) = c_1 a^+_{in,1}(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + c_2 a^+_{in,3}(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + c_3 a^+_{in,2}(\omega)a^+_{in,1}(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3} + e^{i\varphi} c_4 a^+_{in,2}(\omega)a^+_{in,3}(\omega)|0\rangle_{\omega_1}|0\rangle_{\omega_2}|0\rangle_{\omega_3}, \quad (33)$$

where $\varphi = 2arctg\{\frac{2}{\hbar\gamma}\chi_{23}\}$,

that corresponds to CPHASE gate.

IV. Single qubit gates.

If $\omega_1 t = 2\pi m_1,\ \omega_2 t = 2\pi m_2, \omega_3 t = \varphi$, $\chi_{12} = \chi_{23} = 0$ we have using expressions (10-13) for transfer matrix the following expression:

$$PHASE(\varphi) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{-i\varphi} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{-i\varphi} \end{pmatrix}, \quad (35)$$

that corresponds to a single-qubit rotation of our logical qubit around the z axis of the Bloch sphere.

We can also perform the other single-qubit rotation $QET(\theta)$ (that is orthogonal to $PHASE(\varphi)$) by transferring the excitation via waveguide between side cavities of the logical qubit.

V. Logical operations and universality.

The set of operations $\{PHASE(\varphi), QET(\theta), \text{CPHASE}\}$ is a universal one, which means that the presented model is capable of performing arbitrary quantum computations. The universality of this set is actually an encoded one, i.e. it is valid in the Hilbert subspace that corresponds to some logical encoding of qubits. Here we use the pairwise encoding: $|0_L\rangle = |01\rangle, |1_L\rangle = |10\rangle$.

In this encoding any single qubit state $\alpha|0_L\rangle + \beta|1_L\rangle$ is actually stored as an entangled two-qubit state $\alpha|01\rangle + \beta|10\rangle$, that is, the basis state of such a composite qubit is determined by the basis state of the first qubit of a pair.

In this encoding we have the logical rotations around the x and z axes of the Bloch sphere. This allows for performing arbitrary single-qubit transformation and together with a logical CPHASE gate form a universal set of operations.

VI. Summary.

Thus, we have considered two-qubit CPHASE gate operation with arbitrary phase variation on the array of three micro-cavities coupled by cross-Kerr nonlinearity in frequency narrowband regime and single-qubit gates constituting universal set of gates together with two-qubit one. Parameter matching conditions can be introduced for provision of regime with afore desired phase variation. It was shown theoretically in many papers, for example [7,8], that cross-Kerr nonlinearity and hence the efficiency of gates on cross-Kerr nonlinearity can be risen by the use of slow light. Such Kerr nonlinearity could be material or structural, yet only structural one still had yield experimentally this enhancement of nonlinearity [9-13].

Acknowledgment

The research was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, project No. 0671-2020-0065 and by the framework project No. 02.03.2020 No. 00075-02-2020-051 / 1 register No. 78 KBK 01104730290059611.

References

[1] G.J. Milburn, Phys. Rev. Lett. vol. 62, pp. 2124-2127, 1989.

[2] I.L. Chuang and Y. Yamamoto, Phys. Rev. , Vol. A52, pp.3489-3496, 1995.

[3] I-Ch. Hoi, C.M. Wilson, G. Johansson, T. Palomaki, Th.M. Stace, B. Fan, P. Delsing, Phys. Rev. Lett., Vol. 111, 053601, 2013.

[4] J. H. Shapiro, Phys. Rev., Vol. A73, 062305, 2006.

[5] D. Brod, J. Combes and J. Gea-Banacloche, Phys. Rev., Vol. A94, 023833, 2016.

[6] D.J. Brod, and J. Combes, Phys. Rev. Lett., Vol. 117, 080502 (2016).

[7] Chengjie Zhu and Guoxiang Huang, Optics Express, **19**, 23364-23376 (2011).

[8] J.P. Vasco and V. Savona. Slow-Light Frequency Combs and Dissipative Kerr Solitons in Coupled-Cavity Waveguides. Phys. Rev. Applied, **12**, 064065 (2019).

[9] Juntao Li, Liam O'Faolain, Isabella H. Rey, and Thomas F. Krauss. Four-wave mixing in photonic crystal waveguides: slow light enhancement and limitations. Optics Express, **19**, No.5, P.4458-4463 (2011).

[10] Luc Thévenaz, Isabelle Dicaire, Sang-Hoon. Enhancing the light-matter interaction using slow light: towards the concept of dense light. Chin. Proc. of SPIE, **8273**, P.82731D-1-12.

Intellectual Functional Diagnosis Of Large Objects Using Sensor Networks

Gennady Krivoulya
Design Automation Department
KhNURE
Kharkiv, Ukraine
gennady.krivoulya@nure.ua

Vladyslav Shcherbak
Design Automation Department
KhNURE
Kharkiv, Ukraine
vladyslav.shcherbak@nure.ua

***Abstract*— A promising direction in the process of creating expert systems for functional diagnostics is the use of neuro-fuzzy systems. In this article, the following tasks are formulated and solved: carrying out a continuous analysis of the technical state of the diagnostic object in the process of functioning without disrupting functional relations, promptly obtaining information about the technical state of the diagnostic object at an arbitrary moment in time, eliminating the need to use additional stimulating signals for the diagnostic object during the diagnosis process, the ability to predict deviations of the technical state of the diagnostic object from normal in the process of receiving current data from the sensors.**

Keywords*— *expert, diagnostic, system, neuro, fuzzy, diagnosis, object, large, complex, malfunction

I. INTRODUCTION

At present time the advancement of science and technology supposes the use of various large-scale objects in the daily activities of human society. This raises the problem of processing a significant amount of data (Big Data) to determine the state of complex objects in the diagnostic process [1]. Since the data can be unstructured and it have a huge volume and variety, the number of diagnostic parameters (DP) can be in the hundreds or thousands [2][3]. Traditional mathematical methods for large diagnosis object (DO) are inapplicable due to the complexity of describing the states of an object in an explicit form. Therefore, in this situation, one of the possible ways to solve the diagnostic problem is to use intellectual means. Expert knowledge has a decisive role in determining the technical condition of complex objects.

When solving problems of functional diagnostics of complex objects, the main critical factor is the time for making a decision based on production rules in an expert diagnostics system (EDS) for localizing a malfunction [4]. The use of EDS provides decision support in situations for which the diagnostic algorithm is not known and is formed from the initial data in the form of a chain of reasoning (rules). Modern diagnostic objects have difficult dependencies for the formalization of input and output data, therefore it is not always possible to build a strict mathematical model of such objects. To describe the properties of an object, it is advisable to use intelligent models that reproduce the logic of reasoning of the person that makes a decision, the basis of which is the knowledge base (KB) [5].

A significant drawback of expert systems is the significant labor costs required to replenish the knowledge base. Obtaining knowledge from experts and entering it into the knowledge base is a complex process, involving a significant investment of time and money. Therefore, an urgent task is to develop automated methods for replenishing knowledge base for expert systems for functional diagnostics. The solution of this problem is the using of neural networks, which advantage lies in the possibility of transferring the knowledge of the decision maker to the knowledge base of the EDS or automatic replenishment of the knowledge base with data from the sensors of the information system of the DO [6].

If the analysis of the behavior of a complex object is carried out by using a sensor network, then information about the state of the object will received from the sensors at certain intervals. The operator in real time determines the type of malfunction and forms a control action to eliminate the malfunction based on the current readings of the sensors. The initial data for making a decision is the EDS database, which contains models of possible real faults of the object and control actions for their localization or compensation. The information in the database is formed based on the analysis of the readings of the sensors received from the object during a certain interval of observation time in the presence of possible real malfunctions. The large volume of the database and the availability of information in the time coordinate significantly complicates the diagnostic task, therefore, it becomes necessary to structure the diagnostic information in the databases. One of the possible methods for solving this problem is the use of temporal decision trees, which, unlike ordinary decision trees, contain additional information about the time of obtaining information from the corresponding sensors of a technical object. The use of temporal trees can greatly speed up decision-making in an environment where time is a critical factor for decision-making. In this regard, an urgent task is to structure the initial tabular data using temporal decision trees [7].

978-1-7281-9899-6/20/$31.00 ©2020 IEEE

II. PROBLEM DEFINITION

The disadvantage of modern monitoring in the diagnosis of complex technical objects is the inability to determine the initial stage of a failure diagnosis object [7]. The introduction of modern hybrid intelligent technologies in solving diagnostic problems will allow not only to compare the monitored parameters with their reference values, but also to predict the possibility of a malfunction of both individual elements and the object as a whole. A promising direction in the process of creating expert systems for functional diagnostics is the use of neuro-fuzzy systems that combine the advantages of fuzzy expert systems and neural networks. The apparatus of fuzzy logic in the development of a knowledge base (KB), and output mechanisms allows to formalize the procedure of evaluation of technical condition on the basis of unreliable and inaccurate information in the identification of possible malfunctions. In order to form logical conclusions in the form of fuzzy production in hybrid expert diagnosis system (EDS) used in the form of knowledge production with fuzzy linguistic variables, which are presented with the terms of a membership function [8]. EDS (centralized control) at a given time reads the values of the set of parameters from the measuring DO sensors. After that, the expert operator enters the obtained data into the expert system and starts the processing program. Using the operator procedures for processing diagnostic information reduces the efficiency of the EDS and requires additional time. In the presence of automated technical means for storing and collecting information from DO sensors, it becomes possible to automate the procedure for replenishing the knowledge base and track huge volumes of rapidly changing information, to make high-quality and timely decisions when diagnosing complex objects.

III. STATEMENT OF THE PROBLEM

The main purpose of this work is to develop automated methods to replenish the knowledge base of expert systems of functional diagnostics using measuring sensors when monitoring complex objects. In the process of achieving the main goal are formulated and solved the following tasks:

- carrying out continuous analysis of the technical condition of DO during operation without disturbing the functional relationships;
- operative to obtain information about the technical condition of DO at any given time;
- avoid the need for additional stimulus signals for DO in the diagnosis;
- the ability to predict the state of technical deviations from the normal DO in the process of obtaining current data from sensors.

IV. APPLICATION OF TEMPORAL DECISION TREES IN THE INFORMATION PART OF THE EDS

The information part of the expert system ensures the accumulation, storage and transmission of information to other parts of the EDS and implements the end-user interface. The data coming from the sensors is unstructured and requires further processing. It is proposed to use temporal decision trees for preliminary processing and structuring of diagnostic data coming from sensors of a technical object to replenish the knowledge base.

Modeling the behavior of an object considering its faulty states is implemented using a model that describes the structure and behavior of a complex technical object. Such a model is a triple <S, M, R>, where S is a set of variables that describe the state of the system; M – a set of operating modes, including the states "normal" (correct behavior) and «malfunction» (incorrect behavior); R is a set of relations connecting the multitude of variables S, that describes the state of the system and a set of operating modes M.

A table of initial data, which contains all the necessary information to form a tree (Table I), is usually used to construct a temporal decision tree [7].

TABLE I. Table of Initial Data.

Fault	Sensor 1			Sensor 2			Controlling influence	Decision time	The cost
	t_0	t_1	t_2	t_0	t_1	t_2			
F_1	L	L		L	N		U_1	T_1	C_1
F_2	L	N	N	N	H	H	U_2	T_2	C_2
F_i	N	N	N	N	N	N	U_i	T_i	C_i

This table contains m lines, where m is the number of simulated faults; n columns, of which (n – 3) columns correspond to the number of sensors, and the other three contain information about the control action U_i, decision time T_i, and cost C_i of possible consequences for each malfunction. The values of the sensor readings in the form of fuzzy values (N-norm, L-below the norm, H - above the norm) are recorded at the moments of time t_0, t_1, t_2.

Graphically, a temporal decision tree is a weighted directed graph $T_{tmp}=(V_{tmp}, E_{tmp})$, in which v_0 is the root of the tree. All vertices are divided into two classes: V_i - set of internal tree vertices; V_l is the set of outer tree vertices (leaves).

Internal nodes V_i of the tree are weighted by observation, that is, by the pair <a, t_c>, where: a is the name of the attribute; t_c – timestamp. The leaf vertices V_l are weighted by the name or number N_i of the fault, the decision time T_i, the proposed control recovery action U_i, and the cost C_i of the consequences for the fault.

Each arc e of the decision tree is weighted by the condition "attribute [t_c] - attribute value", where "attribute" is the name of the attribute at the vertex from which the arc e originates, "attribute value" is one of the possible values of the "attribute" feature; tc is the time moment at which this check must be carried out, $0 \le t_c < t^*$. When constructing a tree, a restriction is imposed on decreasing time stamps when traversing the tree from the root to the terminal vertex; that is, if time always goes forward, the decision time will generally be different for each situation.

The need to make decisions in real time leads to the fact that the number of trees, that built in accordance with the incoming data, must be equal to the number of samples (analogous to pipelined data processing). Storing decision trees for each time interval requires a significant amount of EDS memory, so averaging is usually used for the input data in order to reduce this cost. However, in this case, information about the current changes in data from the

sensors over a certain period of time can be lost, which is a significant drawback of methods for calculating average values.

The problem of a significant amount of data from a complex DO can be solved by using this data as a training sample for a neuro-fuzzy KB. The EDS considered in this work, along with the use of traditional knowledge in KB, makes it possible to use a neural fuzzy network knowledge base and formalize the above practical problems to achieve the main goal of the work arising during the exploitation of radio electronic equipment.

V. FUZZY MODELING OF TYPES OF FAILURES OF COMPLEX OBJECTS

System of productions, frame structure, semantic networks and logical system are used as the basic models of knowledge representation in intelligent systems. The main difficulties in the design of intelligent diagnostic systems due to the fact that such systems were being developed for poorly formalized subject domains in which knowledge are inaccurate, incomplete, contradictory and variable.

To simulate the reasoning of the expert most appropriate mathematical tool is the language of fuzzy sets, which minimizes the transition from verbal oral qualitative description of the object, which characterizes the human mind to a numerical quantitative estimate of its condition and to formulate on this basis, simple and efficient algorithms. In the fuzzy EDS advanced Boolean logic is used, which represents a set of membership functions and rules to justify

the data. Unlike traditional EDS, which are mainly symbolic inference mechanisms, fuzzy EDS focuses on the numerical processing of source data [9]. For forming the production rules of an expert system, it is necessary to have fuzzy models of possible DO failures.

During research of DO diagnostic state, one of the most complex problems consists in the quantitative description different states of DO subject to faults with emerged in process of exploitation. Solving this problem, it is impossible to correlate many input data with quantitative value. They are often described by qualitative attributes such as “much”, “strong” and so on. Therefore models, which are constructed on numerical estimate input data, are imprecise. Input data also depend on subjective estimate of an expert and enclose uncertainty and ambiguity, which is important to consider in decision-making process.

At present it is known that fuzzy-set theory is useful for problem solution in case, when data are presented in a linguistic expression form and depended on subjective estimate of expert. For DO states evaluation let’s use the variable “Failure” and terms of this variable: {“no”, “light”, “medium”, “strong”, “destructive”}. Each term (value) of linguistic variable defines by fuzzy set. Difference kinds of WSN failures define by classification features, which are presented by fuzzy logic (Table II).

VI. THE STRUCTURE OF THE NEURO-FUZZY KB IN HYBRID EDS

Using a hybrid EDS with a neuro-fuzzy KB for solving problems of diagnosing complex objects expands the capabilities of this class of intelligent systems, allows, with equal computing resources of a computer, to conduct an expert assessment of a larger number of options, increasing the reliability and accuracy of the results obtained [4][6].

The procedure for constructing a hybrid neuro-fuzzy EDS with heterogeneous knowledge for diagnosis under conditions of uncertainty includes the following stages:

- formalization of the subject area (development of a conceptual model);
- selection and adaptation of the diagnostic method;
- a description of the diagnostic model of DO in the form of separate concepts (knowledge) in the knowledge base;
- formation of a knowledge base with a rule base as a control component of the intelligent core;
- description of heterogeneous knowledge in separate subsystems of the hybrid EDS (database, knowledge base, expert knowledge base, graphical database, calculation files, etc.);
- selection of a neural network model and learning rules;
- development of software for the used methods of fuzzy logic;
- distribution of information flows between separate subsystems of the EDS;
- testing of individual EDS subsystems with heterogeneous knowledge and the entire system as a whole.

The main problem in the creation of EDS is the development of the structure of the neural network for the implementation of the neuro-fuzzy KB. This issue is devoted to a lot of scientific publications, which shows the different structures of neural networks for solving the problem. The structure of the neuro-fuzzy network is similar to the structure of a conventional multi-layer neural network with one input layer, one output layer and three hidden layers. Each node of the first layer represents one term with a triangular membership function. In this layer, the values of the membership coefficient are calculated in accordance with the applied fuzzification function for each of the six production inference rules. Layer 2. Antecedents (premises) of fuzzy rules are determined. The output of the node is the degree of compliance of the rule, which is calculated as the product of the input signals. Layer 3. The degree of implementation of production rules is normalized. Layer 4. Conclusions of the rules are formed as values of the weighted components of the output. Layer5. The aggregation of the result obtained according to various rules is carried out. The only neuron in this layer implements the defuzzification operation. Such a neural network makes it possible to identify faults with different degrees of membership [4].

The functional diagnostics algorithm is based on comparing the mathematical model of a specific diagnosed object with its reference and defect-free model, i.e. in checking supplies status parameters are in the range of their changes. Parameter going outside these ranges should indicate the presence of a malfunction in the corresponding object subsystem.

TABLE II. Diagnostic features description

№	Diagnostic features	Fuzzy value of features	Kind of failure	Definition
1	Failure appearance area	significant	hardware failure 	The failure, when the object is failed by reason of hardware failure
		average		
		insignificant	software failure c	The failure, when the object is failed by reason of software failure
2	Nature of parameters variation during the failure	significant (unexpected)	sudden failure 	The failure, which is described by discontinuous variation of object's parameters
		average	progressive failure 	The failure, which appears at object's scaling
		insignificant (gradual)	gradual failure 	The failure, which appears slowly at object's scaling
3	Nature of failure existence on a time	significant (continuous)	hard failure 	The failure, which not ceases before elimination of its causes
		average duration	intermittent failure 	The repeatedly incipient intermittent failure of the same nature
		insignificant (short-time)	fault 	The intermittent or single failure
4	Detection possibility	complex	latent failure 	The failure, which is detected by special diagnostic technique
		average	implicit failure 	The failure, which is produced by another failures and revealed by diagnostic technique
		simple	explicit failure 	The failure, which is detected visually or by using control devices
5	Conditionality by other failures (dependence)	significant	independent failure 	The failure, which isn't conditioned by other failures
		average	implicitly dependent failure 	The failure, which is produced by another failures implicitly
		insignificant	dependent failure 	The failure, which is conditioned by other failures
6	Resilience to break down	complex	unavoidable failure 	The failure, which elimination requires the replacement of misbehaving CS's component
		average	avoidable failure by repair 	The failure, which elimination requires the demounting and repair of misbehaving CS's component
		simple	on-site avoidable failure 	The failure, which elimination is possible without demounting
7	Failure appearance cause	complex	design error failure 	The failure, which is connected with design rule violation
		average	manufacture-error failure 	The failure, which is connected with industrial process violation or WSn repair
		simple	operational failure 	The failure, when uncritical consequence appears
8	Weight of the consequences	significant	critical failure 	The failure, when the danger to person's life and health; and inconsiderable economic losses appear
		average	middle weight failure 	The failure, which can involve an extensive damage but creates small danger to person's life and health
		insignificant	noncritical failure 	The failure, when insignificant consequences appear

In the hybrid neuro-fuzzy EDS, the reference model of the DO is stored in the knowledge base and refined in the process of acquiring new knowledge. The real model is formed in the database environment, and the relationship with the reference model through user queries. The solution of the problem of building an intelligent system for technical diagnostics of the state of DO on the basis of a hybrid EDS made taking into account the features of the external conditions of the EDS environment and the specifics of model adaptation in this environment.

The content, form and algorithms for presenting information in a hybrid ES can be varied depending on the complexity of the simulated situation, the specifics and individual characteristics of the user. The expert user presents expert knowledge of DO diagnostics in the form of sets of examples. The internal form of expert knowledge presentation is a temporal decision tree. A set of examples is described using attributes and contains examples of the same structure, defined by its attributes, which can be connected by logical transitions. In this case, the corresponding decision trees are combined in such a way that another decision tree is added to the terminal vertex of one tree.

VII. The use of EDS for diagnosing large objects

A feature of modern DO for information processing and control is that the diagnostic result depends on the number of input diagnostic parameters (DP) and the corresponding linguistic variables (LV). The initial data at this stage is a list of all possible inputs (diagnostic signs), on which the output depends (the result of diagnostics). Too large number of them will be more difficult diagnosis algorithm, so it is advisable to use only independent diagnostic features. In the manual synthesis of expert knowledge base must be removed from the list of non-essential features that will simplify the diagnosis object model and improve its performance. However, in the case of an automated method for replenishing a knowledge base, the number of input variables is determined by the number of EDS sensors.

For an example of the EDS functioning, we consider that it is possible to measure numerical values for 24 diagnostic parameters (DP_1, ..., DP_{24}). The sensor readings are obtained at discrete times t_0, t_1, t_2,…, t_i. The time interval (t_{i+1} - t_i) between two adjacent measurements is selected taking into account the rate of change of diagnostic parameters. All 24 characteristics will play the role of diagnostic parameters in the intelligent diagnosis process.

In evaluating the input DP (determining the diagnostic parameters) in the simplest case, it is advisable to limit ourselves to the three levels of gradation (three terms): L (low, below normal), A (average, norm), H (high, above normal).

Consider the membership function of an arbitrary input variable DP_i, which corresponds to some diagnostic feature, for example, "soil temperature" (Figure 1). The table for this example contains the parameters DP_i.

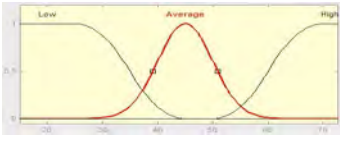


Fig. 1. Membership function DPi

To implement the diagnostic algorithms, we use the MatLab and adaptive-network-based fuzzy inference systems (ANFIS) is used to solve problems related to parameter identification. ANFIS is basically a graphical network representation of Sugeno-type fuzzy systems endowed with the neural learning capabilities. The network is comprised of nodes with specific functions collected in layers. ANFIS is able to construct a network realization of IF/THEN rules. All computations can be presented in a diagram form. ANFIS normally has 5 layers of neurons of which neurons in the same layer are of the same function family.As a result of computer modeling, the structure of a neural fuzzy network for 24 inputs was obtained, which allows diagnosing possible DO failures. [9]

VIII. Conclusions

The time for making a decision on the localization of a fault is the most important critical factor in determining the technical state of complex objects. Using a hybrid expert diagnostic system with a neuro-fuzzy network knowledge base provides decision support in situations for which the diagnostic algorithm is not known and is formed from the initial data in the form of production rules. To automate the process of accumulating knowledge in an expert system, it is advisable to use object sensors, with which the values of diagnostic parameters are measured. The initial data is structured using temporal decision trees. The need to make decisions in real time leads to the fact that the number of trees corresponding to the incoming data is equal to the number of samples during the observation time. The problem of a significant amount of data in determining the state of a complex technical object is solved by using this data as a training sample for a neuro-fuzzy knowledge base.

References

[1] Asside Christian Djedouboum, Ad Adamou Abba Ari, Abdelhak MouradGueroui, Alidou Mohamadou and Zibouda Aliouat. Big Data Collection in Large-Scale Wireless Sensor Networks. Sensors 2018, 18, 4474, pp. 1-34

[2] Y. Liu, X. Mao, Y. He, K. Liu, W. Gong, and J. Wang, “Citysee: not only a wireless sensor network,” *IEEE Network*, vol. 27, no.5, pp. 42–47, 2013.

[3] Y. Qu, W. Han, L. Fu et al., “LAINet - A wireless sensor network for coniferous forest leaf area index measurement: Wireless Communications and Mobile Computing 17 Design, algorithm and validation,” *Computers and Electronics in Agriculture*, vol. 108, pp. 200–208, 2014.

[4] Sensor Networks. In Computational Intelligence in Sensor Networks; Springer: Berlin, Germany, 2019; pp. 215–248.

[5] G. Krivoulya, A. Shkil, D. Kucherenko, A. Lipchansky, Ye. Sheremet. Expert evaluation model of the computer system diagnostic features // EWDTS'2014: Proceeding of international conf., 26-29 September, 2014. – Kiev, Ukraine, 2014. – pp. 286-289.

[6] Polkovnikova N. A., Kureychik V. M. Ob intellektual'nom analize baz dannykh dlya ekspertnoy sistemy // Informatika, vychislitel'naya tekhnika i inzhenernoye obrazovaniye. - 2013 - № 2 (13). - S. 39-50.K. Elissa, “Title of paper if known,” unpublished.

[7] Krivulya G. F., I. V. Vlasov, O. A. Pavlov. Operativnoye funktsional'noye diagnostirovaniye tekhnicheskikh ob"yektov s primeneniyem temporal'nykh derev'yev resheniy. Sb. nauchnykh trudov konferentsii «Intellektual'nyye sistemy prinyatiya resheniy i problemy vychislitel'nogo intellekta». Yevpatoriya - 2013 S.193-195.

[8] C. Loganathan, K. V. Girija. Hybrid Learning For Adaptive Neuro Fuzzy Inference System. International Journal Of Engineering And Science Vol. 2, Issue 11 (April 2013), pp 06-13.

[9] Krivoulya G. Expert diagnosis of computer systems using neuro-fuzzy knowledge base/ Krivoulya G., A. Lipchansky, Ye. Sheremet EWDTS'2016: Proceding. of international conference., 15-17 September, 2016. – Erevan, – P. 619-622

AUTHOR INDEX

Camera-ready was prepared by Chumachenko S.
Approved for publication: 03.09.2020. Format 60x841/8.
Relative printer's sheets: . Circulation: 100 copies.
Published by SPD FL Stepanov V.V.
Ukraine, 61168, Kharkov, Ak. Pavlova st., 311