Journal of Physics: Conference Series

PAPER • OPEN ACCESS

The LHCb Grid Simulation: Proof of Concept

To cite this article: M Hushchyn *et al* 2017 *J. Phys.: Conf. Ser.* **898** 052020

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 052020 doi:10.1088/1742-6596/898/5/052020

The LHCb Grid Simulation: Proof of Concept

M Hushchyn[1,2], A Ustyuzhanin[1,3], K Arzymatov[2], S Roiser[4] and A Baranov[1]

[1]Yandex School of Data Analysis, Moscow, Russia
[2]Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[3]National Research University Higher School of Economics, Moscow, Russia
[4]CERN, Geneva, Switzerland

E-mail: mikhail.hushchyn@cern.ch

Abstract. The Worldwide LHC Computing Grid provides access to data and computational resources to analyze it for researchers with different geographical locations. The grid has a hierarchical topology with multiple sites distributed over the world with varying number of CPUs, amount of disk storage and connection bandwidth. Job scheduling and data distribution strategy are key elements of grid performance. Optimization of algorithms for those tasks requires their testing on real grid which is hard to achieve. Having a grid simulator might simplify this task and therefore lead to more optimal scheduling and data placement algorithms. In this paper we demonstrate a grid simulator for the LHCb distributed computing software.

1. Introduction

The Worldwide LHC Computing Grid (WLCG) [1] is a global collaboration of computing centers which provides computing resources to store, distribute and analyse the data produced by the Large Hadron Collider (LHC) as well as corresponding simulated data. This makes the resources and the data available to all partners with different physical locations.

According to the WLCG public site [1], the WLCG has a hierarchical topology and consists of three layers, i.e. tiers. The Tier 0 is the CERN Data Center, which is located in Geneva, Switzerland and also at the Wigner Research Center for Physics in Budapest, Hungary. Tier 1s are large computer centers with sufficient storage capacity. There are fourteen Tier 1s around the world. Tier 2s are typically universities and other scientific institutes. There are currently about 150 Tier 2s. The Tiers resources are presented on the WLCG REBUS [2].

CERN Data Center and Wigner Research Center for Physics are connected by three 100 Gb/s lines. These lines create a combined Tier 0. CERN connects every Tier 1 around the world using a dedicated high-bandwidth network, the LHC Optical Private Network (LHCOPN) [3]. This consists of optical-fibre lines of at least 10 Gb/s, spanning oceans and continents. WLCG also utilizes the LHC Open Network Environment (LHCONE) [4] which complements the LHCOPN.

The WLCG resources are shared and managed by the LHC experiments and the participating computer centers. The LHCbDirac [5] interware provides access to the LHCb resources and handles all the distributed activities of LHCb. These activities are Monte-Carlo simulation, data processing, data management, resources management and monitoring, accounting for user and production jobs, etc..

A simple simulation of the LHCb grid is demonstrated in this study. Two job scheduler strategies are simulated to explore their impact on the grid performance. The goal of the study

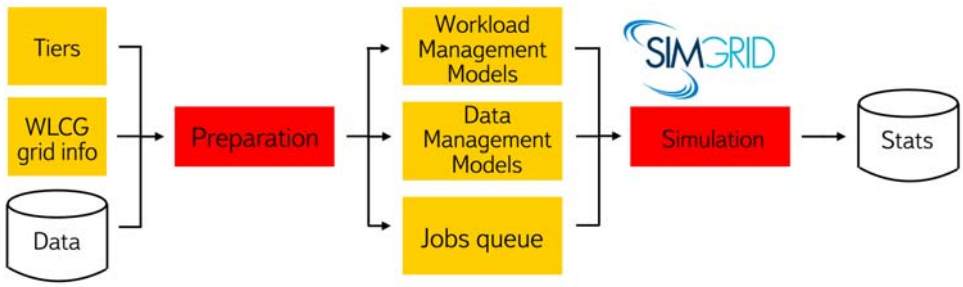

Figure 1. The simulation workflow.

is to prove the concept that the simulation can be used to select the optimal strategy of using the available distributed computing resources.

2. The LHCb Computing Model

As described in [6], LHCb uses a sophisticated two-level trigger system for selecting interesting collisions coming from LHC. These selected collisions are recorded to RAW files by the LHCb online system. Then, the RAW files are transferred to the Tier 0 data storage system and replicated to one of 7 LHCb Tier 1s [2] on custodial storage as well.

After the replication, the RAW data is reconstructed at CERN and on the Tier 1s. Output data (SDST files) is uploaded to custodial storage at the site where it is produced.

The reconstructed data goes through very severe selections which correspond to different types of physics analysis. This stage is called stripping and produces mostly DST files. These files are small and it is convenient to merge them into files of 5 GB. The merged files are replicated according to the baseline policy: 3 replicas on disk storage at Tier 1s and one at CERN, 1 archive replica on tape storage at CERN and one at a Tier 1.

Monte-Carlo simulation of real collisions is highly CPU consuming and performed on all available sites. Output data is uploaded to Tier 1s data storage. Simulation data goes through reconstruction and stripping as well as real data. DST files are replicated as follows: 2 replicas on disk storage at Tier 1s and one at CERN, one archive replica on tape storage at CERN or a Tier 1.

Physics analyses are performed using the DST files at tiers where these files are replicated.

Data is managed by the LHCbDirac Data Management System (DMS) [5, 6]. Data Management includes data replication on disk and tape storage systems, replica removal to free disk space for newer data and data checks.

Data processing jobs are managed by the LHCbDirac Workload Management System (WMS) [5, 7]. Workload Management activities include job submission and monitoring, job scheduling, etc..

3. SimGrid Framework

SimGrid [8] is a framework to simulate distributed computer systems: grids, clusters, IaaS clouds, high performance computing, volunteer computing and peer-to-peer systems. This framework can be used to develop new algorithms or to debug real distributed applications.

A SimGrid simulation contains the following components:

- **Application:** a distributed algorithm described in SimGrid APIs.
- **Virtual Platform:** description of a given distributed system (machines, disks, links, clusters, etc.) in XML format. SimGrid supports dynamic scenarios where the platform parameters can be changed during a simulation.

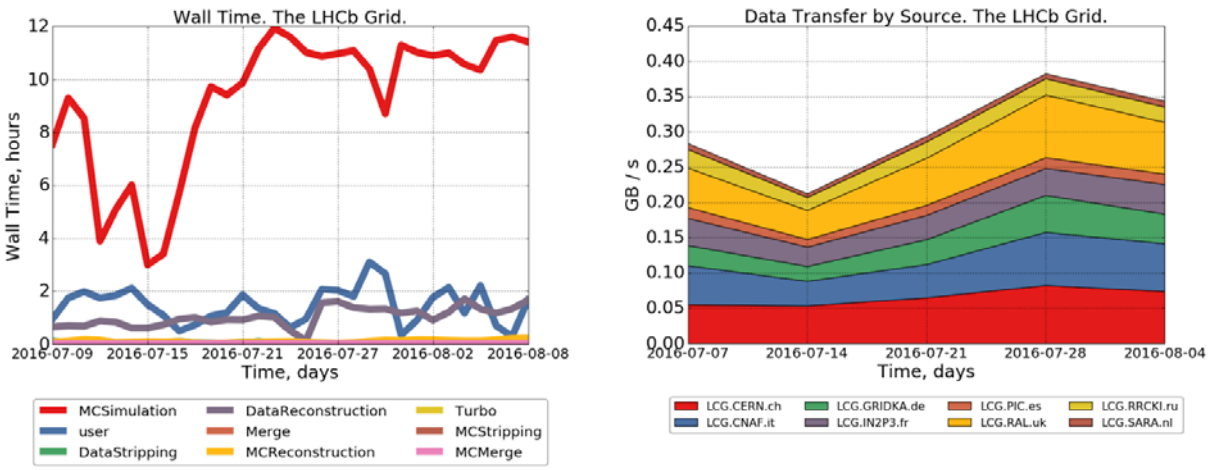

Figure 2. The LHCb Grid operation. Left: average wall time for the job types. Right: average data transfers.

- **Deployment Description:** description of how the application should be deployed on the virtual platform. For example, which process is located on which host.
- **Platform Models:** description of how the virtual platform reacts to actions of the application. These models are already included in SimGrid. Thus, it is needed only to select one.

4. Simulation Workflow

The simplified LHCb Grid model is simulated. This model contains Tier 0 at CERN and those seven Tier 1s [2] which are used by LHCb with corresponding resources of CPU, disk, and tape storage. Tiers are connected by the LHCOPN [3] network.

The following data processing job types are simulated: Data Reconstruction, Data Stripping, Merge, MC Reconstruction, MC Stripping and MC Merge. The Monte-Carlo data is generated by MC Simulation jobs. Turbo jobs [9] process the data with higher output rate. User jobs correspond to physics analysis. The fraction of every job type contribution, average CPU time, and input/output data size were estimated using the LHCb Dirac Web Portal[1] and then were used for the job queue generation.

Two job schedulers with the "pull" paradigm are considered. The first one is Simple Model: a job is sent to the first available site, regardless of the input data location. The second model is Data Availability Model (DAM). In this case, a job is sent to the first available site which has the required input data. Both simulations assume the replication strategy provided by the LHCbDirac DMS [6].

The simulation is performed using SimGrid [8]. The whole simulation workflow is presented in figure 1.

5. Simulation Results

The job wall time and the average data transfer rate for the real grid are shown in figure 2. Data presented in the figure is provided by the LHCb Dirac Web Portal. The simulation results are demonstrated in figure 3. The average job wall time for the DAM model and real grid are compatible. On the other hand, the results for Simple Model differ significantly. The reason for

[1] The LHCb Dirac Web Portal `http://lhcb-portal-dirac.cern.ch/DIRAC/`

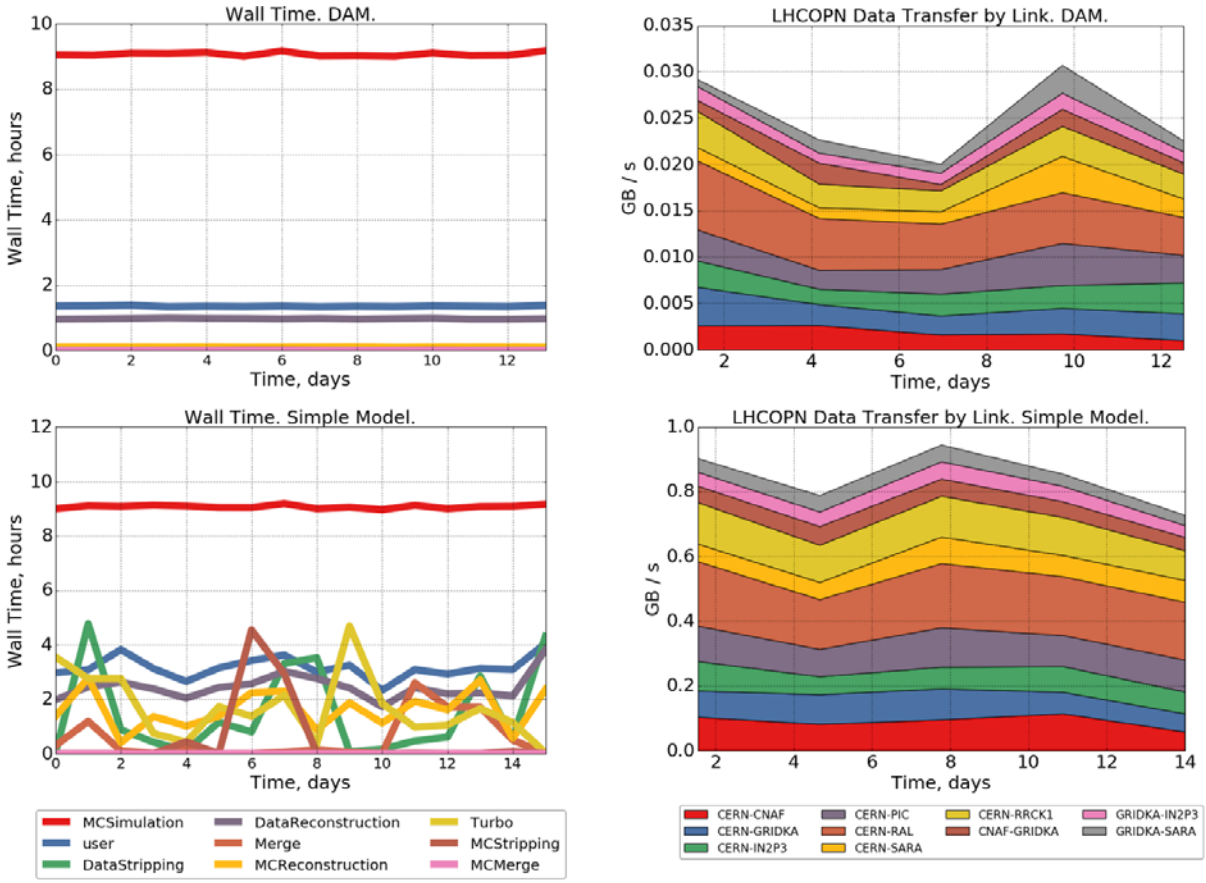

Figure 3. The simulation results. Left column: average wall time for the job types. Right column: average data transfers.

this inconsistency is that jobs should wait for the input data transfer from the other sites. This leads to the increased wall time. The figure shows that the average data transfer rate is similar for the LHCb Grid and DAM model. The transfer rate for Simple Model is much higher.

6. Conclusion

The simple example demonstrates that simulation allows to estimate the influence of different strategies on the LHCb Grid performance and choose the best one. However, this approach needs further improvement and development to increase the precision of results.

7. Acknowledgments

The authors are grateful to Fedor Ratnikov for his helpful comments during the work on this study and help with preparing this paper.

References

[1] The Worldwide LHC Computing Grid `http://wlcg-public.web.cern.ch`
[2] The WLCG Resource, Balance and Usage `https://wlcg-rebus.cern.ch/apps/topology/`
[3] The LHC Optical Private Network `http://lhcopn.web.cern.ch/lhcopn/`
[4] The LHC Open Network Environment `http://lhcone.web.cern.ch`
[5] Stagni F etc. 2012 LHCbDirac: distributed computing in LHCb *J. Phys.: Conf. Series* **396** 032104

[6] Baud J P, Charpentier P, Ciba K, Graciani E, Lanciotti E, Mth1 Z, Remenska D and Santana R 2012 The LHCb Data Management System *J. Phys.: Conf. Series* **396** 032023
[7] Casajus A 2012 Status of the DIRAC Project *J. Phys.: Conf. Series* **396** 032107
[8] Casanova H, Giersch A, Legrand A, Quinson M and Suter F 2014 Versatile, Scalable, and Accurate Simulation of Distributed Applications and Platforms *Journal of Parallel and Distributed Computing* **74** 2899-2917
[9] Benson S, Gligorov V, Vesterinen M A and Williams J M 2015 The LHCb Turbo Stream *J. Phys.: Conf. Series* **664** 082004