Simulation Analysis of CMS Data Replication and Production Activities

Ciprian Dobre, Valentin Cristea
Department of Computer Science and Engineering
University POLITEHNICA
Bucharest, Romania
E-mails: {ciprian.dobre, valentin.cristea}@cs.pub.ro

Abstract— **The scale, complexity and worldwide geographical spread of the LHC computing and data analysis problems are unprecedented in scientific research. The complexity of processing and accessing this data is increased substantially by the size and global span of the major experiments, combined with the limited wide area network bandwidth available. We present the latest generation of the MONARC (MOdels of Networked Analysis at Regional Centers) simulation framework, as a design and modeling tool for large scale distributed systems applied to HEP experiments. We present simulation experiments designed to evaluate the capabilities of the current real-world distributed infrastructure to support existing physics analysis processes and the means by which the experiments bands together to meet the technical challenges posed by the storage, access and computing requirements of LHC data analysis within the CMS experiment.**

Keywords: Modeling and simulation; evaluation; large scale distributed systems; LHC experiments; CMS.

I. INTRODUCTION

Modeling and simulation were seen for a long time as viable solutions to develop new algorithms and technologies and to enable the enhancement of large-scale distributed systems, where analytical validations are prohibited by the scale of the encountered problems. The use of discrete-event simulators in the design and development of large scale distributed systems is appealing due to their efficiency and scalability.

The Large Hadron Collider (LHC) is a giant particle accelerator consisting of a circular tunnel with a circumference of 27 km (the largest in the world), around which beams of protons and anti-protons (and heavy ions such as lead nuclei) are accelerated in opposite directions to nearly the speed of light [3]. At four points on the ring of the accelerator, the beams of particle and anti-particles cross and collide with each other at extremely high energies, close to the energies of the first split seconds after the Big Bang, to produce other kind of particles. The experiments building each detector are ALICE, ATLAS, CMS and LHCb, and each is designed to study a different area of particle physics.

The scientific wealth of the experiments presents new problems in data access, processing and distribution, and collaboration across national and international networks, on a scale unprecedented in the history of science. The information technology challenges are introduced by the need to provide rapid access to data subsets drawn from the massive data stores. Approximately 10-14 Petabytes of data need to be handled and store, and it is expected that the volume of the data will increase in the following years.

The size of the LHC experiments and the unprecedented scale of data resulted in the need to look at resources outside of CERN. From the beginning it was clear that to process all the data centrally at CERN was not a practical or viable solution. Instead, physicists from all over the world offered their own existing resources to be used in the experiments. Today all LHC experiments are embracing the hierarchical distribution model, according to which facilities from all around the world are putting together resources in order to provide the necessary computing power and data storage space needed for the experiments [4]. According to this model the system is composed of an assembly of distributed computing resources, concentrated in a hierarchy of centers called Tiers, where Tier0 is CERN, Tier1s are the major computing centers which provide a safe data storage, likely in the form of a mass storage system (MSS), and Tier2s are smaller regional computing centers.

As the LHC experiments are currently well underway physicists are interested in evaluating the capability of the currently deployed (networking and computational) resources to handle the large amount of data and processing requirements. The difficulty in simulating the running conditions of the physics experiments comes from the large amount of resources involved in the analysis procedures, as envisioned by the computing models [5].

The evaluation of such complex simulations is hard to accomplish using existing simulators. SimGrid [6] is a simulation toolkit that provides core functionalities for the evaluation of scheduling algorithms in distributed applications in a heterogeneous, computational Grid environment. It aims at providing the right model and level of abstraction for studying Grid-based scheduling algorithms and generates correct and accurate simulation results. GridSim [7] is a grid simulation toolkit developed to investigate effective resource allocation techniques based on computational economy. OptorSim [8] is a Data Grid simulator designed specifically for testing optimization techniques to access data in Grid environments. OptorSim adopts a Grid structure based on a simplification of the architecture proposed by the EU DataGrid project. Given a replication algorithm and a Grid configuration as an input, it runs various activities over its resources.

Such simulators were developed for particular classes of experiments. They all support, to some extent, the simulation of data transfer and replication techniques. However, they do

not present general models that allow the evaluation of replication in the wider context of different architectures encountered in case of distributed systems. The simulation instruments tend to narrow the range of simulation scenarios to specific subjects, such as scheduling or data replication.

MONARC 2, using highly advanced technologies to cope with the simulation of large amount of resources and applications, such as the ones described in the computing model of the CMS experiments [5], is able to successfully test the running conditions of the LHC experiments.

In this paper we present experiments designed to evaluate the capability of the evaluate the capabilities of the current real-world distributed infrastructure to support existing physics analysis processes and the means by which the experiments bands together to meet the technical challenges posed by the storage, access and computing requirements of LHC data analysis within the CMS experiment.

The rest of the paper is structured as follows. Section 2 gives a analysis of related work. In Section 3 we present the MONARC simulation model. Section 4 presents implementation details and results of the experiments designed to evaluate the running conditions of the CMS experiment. Finally, in Section 5 we give conclusions and present future work.

II. MONARC Simulation Framework

MONARC 2 is built based on a process oriented approach for discrete event simulation, which is well suited to describe concurrent running programs, network traffic as well as all the stochastic arrival patterns, specific for such type of simulation [1][2]. Threaded objects or "Active Objects" (having an execution thread, program counter, stack...) allow a natural way to map the specific behavior of distributed data processing into the simulation program.

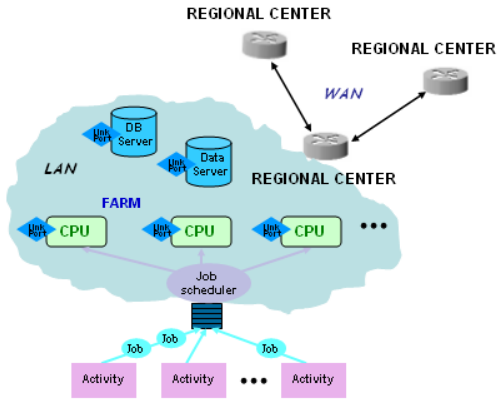

Figure 1. The Regional center model.

In order to provide a realistic simulation, all the components of the system and their interactions were abstracted. The chosen model is equivalent to the simulated system in all the important aspects. A first set of components was created for describing the physical resources of the distributed system under simulation. The largest one is the regional center (see Figure 1), which contains a farm of processing nodes (CPU units), database servers and mass storage units, as well as one or more local and wide area networks. Another set of components model the behavior of the applications and their interaction with users. Such components are the "Users" or "Activity" objects which are used to generate data processing jobs based on different scenarios. The job is another basic component, simulated with the aid of an active object, and scheduled for execution on a CPU unit by a "Job Scheduler" object.

With this structure it is possible to build a wide range of models, from the very centralized to the distributed system models, with an almost arbitrary level of complexity (multiple regional centers, each with different hardware configuration and possibly different sets of replicated data).

III. CASE STUDIES FOR THE LHC EXPERIMENTS

The hierarchical distribution architecture is well mapped on the proposed simulation model. The simulation model allows the simulation of this type of organization. In the computing model of CMS [5] the collections of processing nodes, data warehouses and networking entities are organized in what is called regional centers. The network simulation model allows these regional centers to be connected in arbitrary architectures, including the hierarchical model proposed by the physics experiments. Special designed job models designed to imitate the behaviour of the running LHC conditions are also integrated into the simulation model. These elements allow the easy construction of simulation experiments designed to test the running conditions of the LHC experiments, as envisioned in the computing models.

The general concept developed by the CMS experiments is a hierarchy of distributed Regional Centers working in close coordination with the main center at CERN. This simulation study follows this concept and describes several major activities; mainly the data transfer on WAN between the T0 (CERN) and a number of several T1 Regional Centers. The topology describing the connectivity of the Regional Centers is presented in Figure 2.

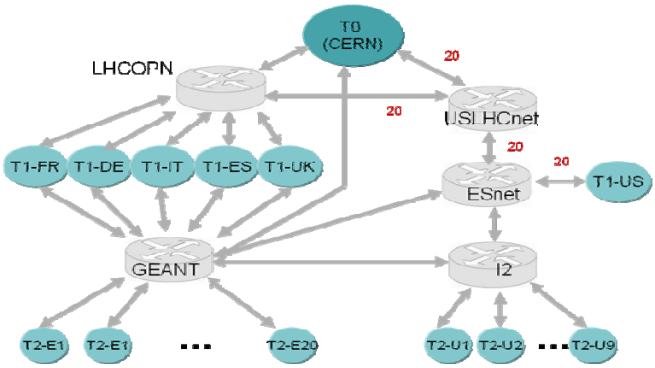

Figure 2. The simulation scenario resembling the CMS computing model.

We assume that the five T1 Regional Centers in Europe are connected independently, by two networks: GEANT (external to CERN) and LHCOPN (within CERN). In a simplified model this can be approximated with two "mega-

routers" in which each T1 regional center is connected through a link. We also consider several transatlantic links connecting T0 with the regional centers in US, through the USLHCnet and ESnet "mega-routers".

We first executed a series of simulation experiments designed ot evaluate the function of the MONARC's model, its capacity to handle the scenario and conditions of the running CMS experiments. These experiments are based on queuing models to evaluate the experiments.

In these experiments events produced in the LHC detector are transferred and processed at different regional centers. We first evaluated the behavior of the database as more events are concurrently served (*the first series of experiments*). Next we evaluated the behavior of the network as more events are concurrently transferred (*the second series of experiments*), and the use of uniform and non-uniform approaches to transfer the data. The analytical results allowed us to compare them against the obtained ones, thus validate the model and experiment.

In these experiments we considered a scenario consisting of several regional centers (see Figure 3) – a simplified version of the actual CMS experiments. We also evaluated the capability of the MONARC simulator to handle large scale experiments. As such, we successfully simulated for example 10000 concurrent jobs, each concurrently transferring and processing 100 events, with concurrent processors and databases.

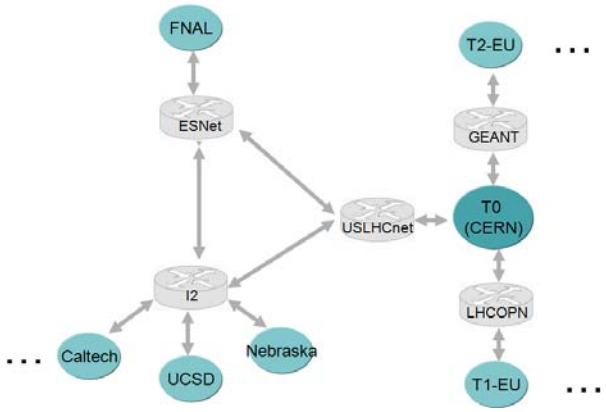

Figure 3. The simplified scenario.

In the first series of experiments we considered several jobs concurrently running at Caltech. Each job reads and processes 100 events from a database situated at UCSD (see Figure 3). These experiments were designed to evaluate the simulation model. The database at UCSD can serve data at a speed of 100 Mbps. The Caltech center contains 10 processing cores. The size of an event is approximately 400 KB (we actually used a normal distribution, so sizes are values in the range 300-500 KB). The processing of an event requires from 1.4 to 2.4 seconds (depending on the size).

Each job transfers an event, and processes it locally. The algorithm continues for the other 100 events. Under normal conditions, in an experiment involving 20 such jobs running concurrently, the experiment would show that approximately 800 MB of data are transferred, and the average CPU load at Caltech is around 80%. This is consistent with the results obtained in the simulation conducted for this case (see Figure 4).

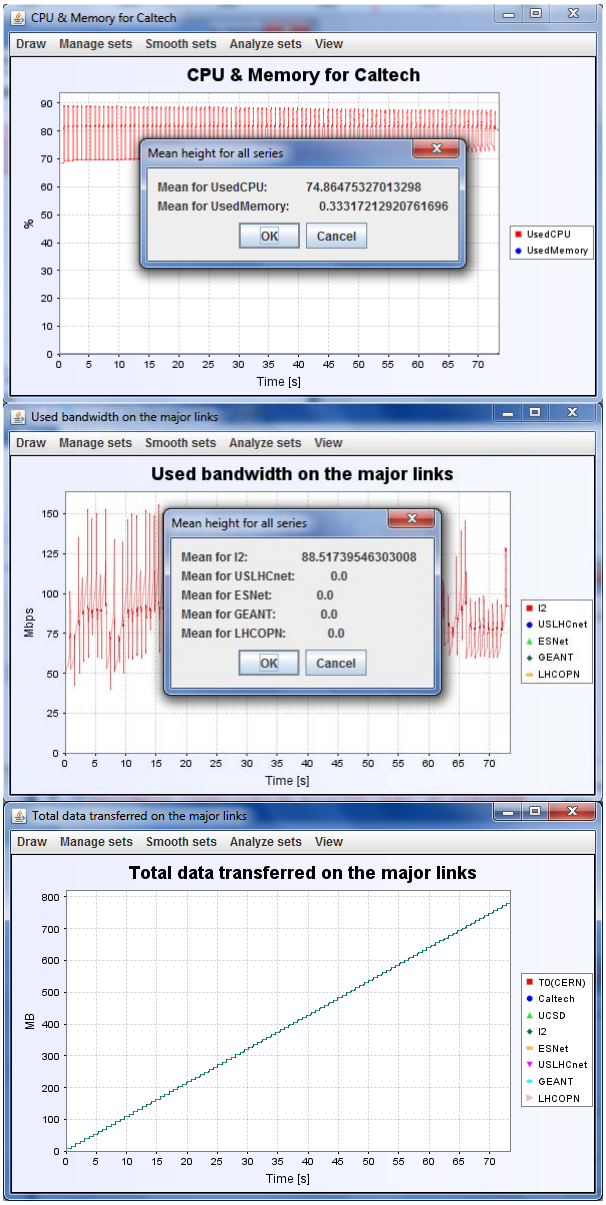

Figure 4. Validation results showing the average CPU load at Caltech, the throughput, and the total data transferred.

We continued with experiments by graduately increasing the number of jobs running concurrently within Caltech. The other parameters were kept constant. We were interested in the capability of the database server at UCSD to handle the large number of requests. As we increased the number of jobs, the load on the database also increased, up to a point where the delays started to affect the performance of the jobs. For example, Figure 5 (left) shows the results for the CPU usage obtained in these experiments. The horizontal axis represents the number of concurrent jobs being used in the experiment. On the vertical axis the values represent the average CPU usage registered at Caltech. In the beginning,

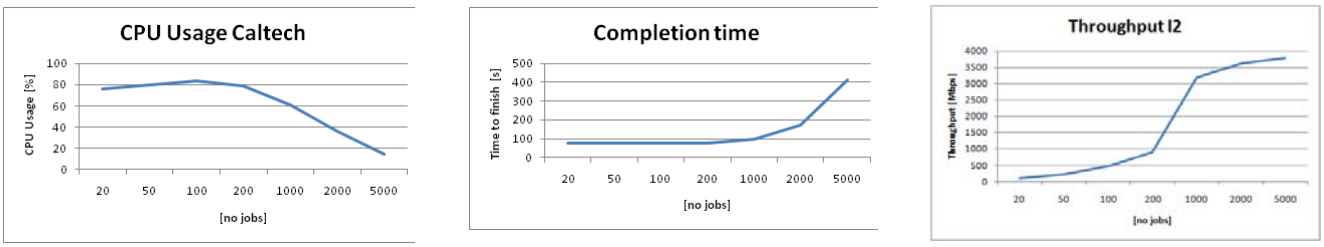

Figure 5. Results showing the relation between the database saturation and the effect on the processor (left), time to finish (center), and throughput on the Internet2 (right).

Figure 6. Results showing the relation between the network saturation and the effect on the processor (left), time to finish (center), and throughput on the Internet2 (right).

as more jobs run and process events concurrently, the CPU usage increases as expected. At around 100 jobs the database reaches an internal bottleneck and starts serving events slower. Starting this point the time needed for the events to be transferred locally increases, so the overall CPU usage decreases. The results are also observed in the time needed to complete the simulation increases as more jobs rush concurrently to get the data from the database server at UCSD.

We next continued evaluating the network conditions. These experiments involved 10 databases located at UCSD, all capable to serve the events. This allow us to relieve the load on the database servers. But, as expected, in this case the network capacity becomes the limit. We executed a series of experiments by varying the number of concurrent jobs. The results in Figure 6 show the relation between the network saturation and the effect on the CPU usage, completion time and the throughput on the Internet 2 link. As the time to transfer an event increases with each experiment, due to the network link becoming a bottleneck, the CPU usage decreases. These results are sustained also by the completion and throughput values (center and right).

A comparison between the first series of experiments and the second one, for the same case of 5000 concurrent jobs running at Caltech and requesting events from the database(s) at UCSD, is presented in Figure 7.

In the third series of experiments we evaluated different solutions to distribute the events. In these experiments the jobs run at Caltech and transfer events from three external regional centers (FNAL, CERN and UCSD) before processing them. In the first experiments we considered a uniform distribution of events. For that each job requests a uniform event identifier normally distributed over the three considered centers. In the second case we considered three distinct sets of jobs that take data only from one distinct regional center from the three considered.

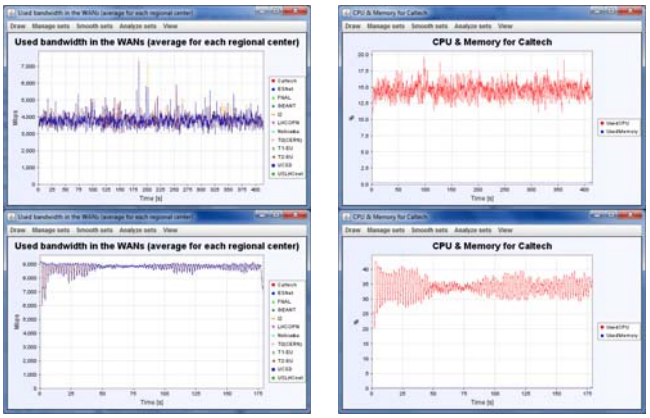

Figure 7. Results obtained in the first series of experiments (up) and the second one (down).

A comparison of the results obtained in these cases, for the throughput, is presented in Figure 8. Figure 9 shows the results presented 10000 jobs 100 events

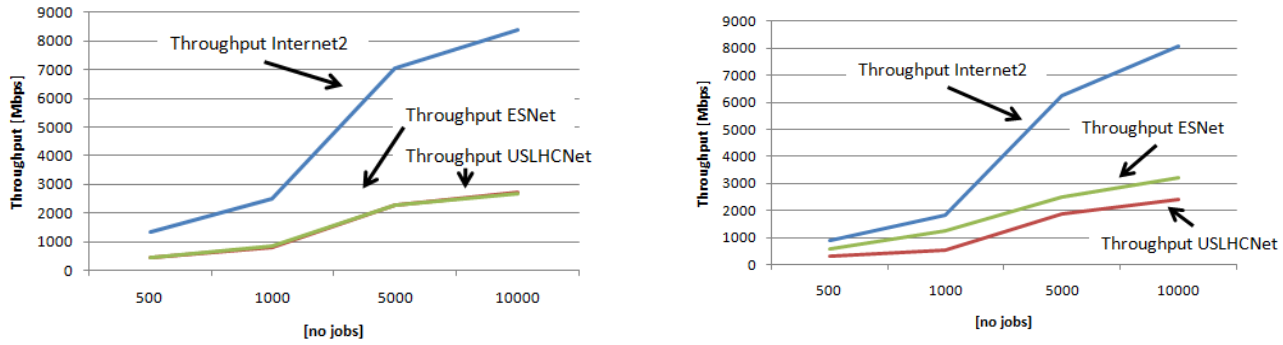

Figure 8. Results obtained for the throughput in case of uniform (left) and non-uniform (right) distributions.

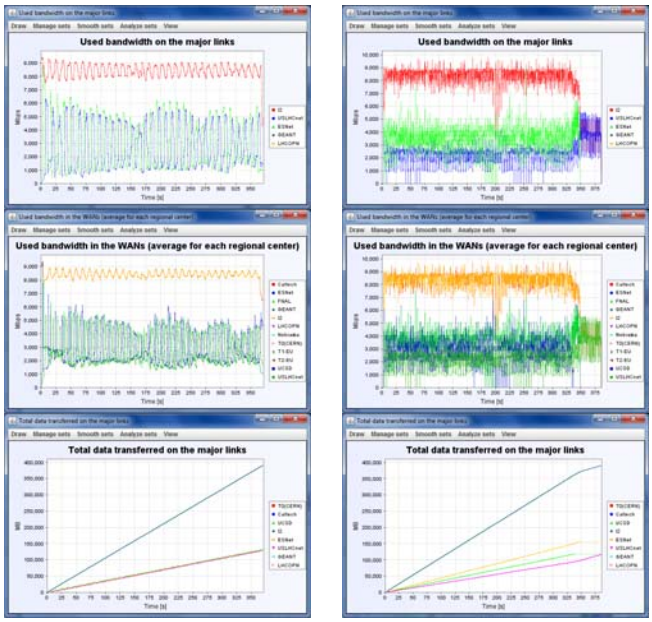

Figure 9. Comparison between experiments using uniform (left) and non-uniform (right) distributions of events.

IV. A SIMULATION STUDY FOR T0/T1 DATA REPLICATION & PRODUCTION ACTIVITIES

After the validation of the model we proceeded with the translation of the topology presented in Figure 2. The end-result simulation model is presented in Figure 10. It uses the entities available within the MONARC simulator.

The values on the links represent the available bandwidth (in Gbps). For a better representation, this topology was simplified and the Tier2 centers were purposely ignored. Using this topology we simulated a number of Activities specific for Physics Data Production, as follows.

RAW Data Replication. From the experiment we assumed a mean rate of recording raw data equal to 200 MB/s. This information is stored in 2GB (normal distributed with 10% sd) data files. These files are replicated in a round robin manner to all 6 T1 regional centers. (The first file is sent to T1-FR, the second to T1-DE…).

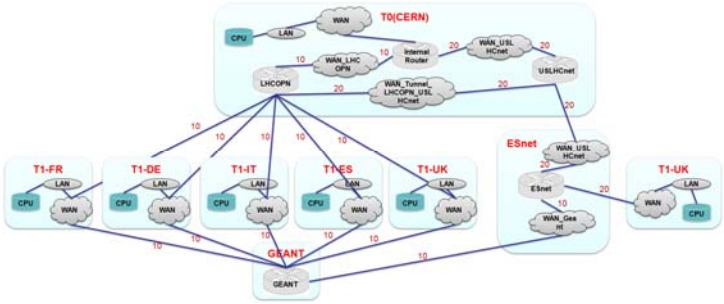

Figure 10. The configuration used in the CMS modeling experiments.

Production and DST distribution. At T0 all raw data are processed and DST files are generated. The DST files are 10 times smaller in size than the RAW files. We considered again a normal distribution (sd 10%). The DST files created at T0 are sent to all T1 centers.

Re-production and new DST distribution. After a certain time the RAW data in each T1 center is re-processed and new DST data is created. Each T1 center will reprocesses 1/6 of the RAW data. The DST data generated at each regional center are sent to all others.

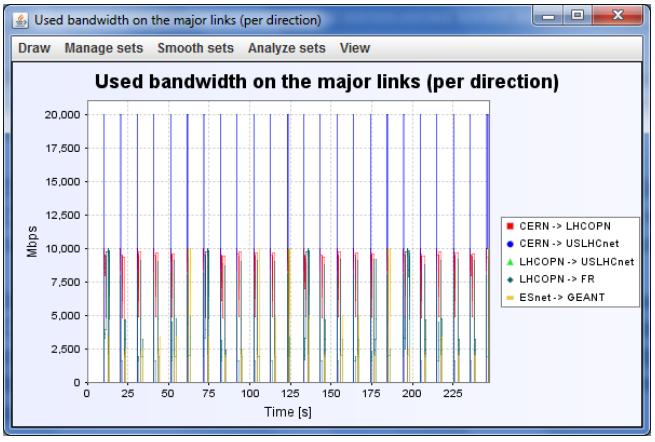

Figure 11. The bandwidth used in the major networks.

We evaluated all three activities (RAW Data Replication, Production and DST distribution and Re-production and new DST distribution) running in parallel. The conditions are the same, as illustrated in Figure 2, for the links connecting the Regional Centers. We first executed a series

of experiments for calibration. We considered a limited set of rounds to evaluate the correctness of the proposed experimental scenario. Figure 11 shows the bandwidth usage on the major networks within the scenario. The network usage occasionally fills the available networks. The same conclusions are drawn from the results presented in Figure 12.

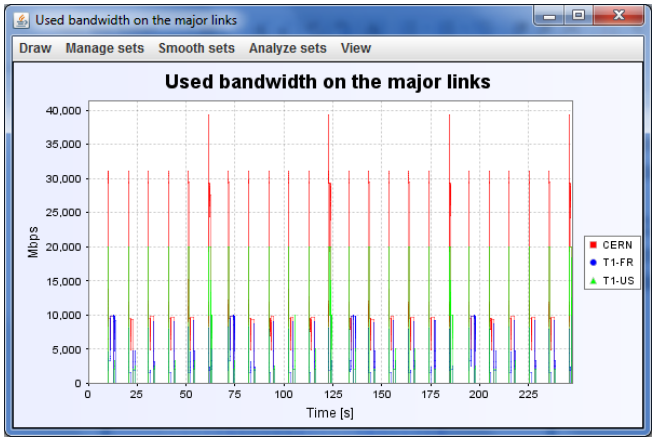

Figure 12. The bandwidth used in the major links.

V. CONCLUSIONS

Large scale distributed systems are currently progressing from operational infrastructures to environments providing many "modern" capabilities. As the LHC experiments are currently well underway physicists are interested in evaluating the capability of the currently deployed (networking and computational) resources to handle the large amount of data and processing requirements.

Simulation is an attractive alternative to evaluating such solutions. However, the evaluation of such complex simulations is hard to accomplish using existing simulators. Previous simulators were developed for particular classes of experiments. They do not present general models that allow the evaluation of replication in the wider context of different architectures encountered in case of distributed systems. The simulation instruments tend to narrow the range of simulation scenarios to specific subjects, such as scheduling or data replication.

In this paper we proposed a series of experiments designed to evaluate using MONARC the capabilities of the current real-world distributed infrastructures at CERN to sustain the LHC experiments. These experiments are designed to evaluate the existing physics analysis processes and the means by which the experiments bands together to meet the technical challenges posed by the storage, access and computing requirements of LHC data analysis within the CMS experiment. The results demonstrate the capability to correctly model solutions for large scale distributed systems.

ACKNOWLEDGMENT

The research presented in this paper is supported by national project: "TRANSYS – Models and Techniques for Traffic Optimizing in Urban Environments", Contract No. 4/28.07.2010, Project CNCSIS-PN-II-RU-PD ID: 238. The work has been co-funded by national project ``DEPSYS - Models and Techniques for ensuring reliability, safety, availability and security of Large Scale Distributes Systems", Project CNCSIS-IDEI ID: 1710. The work has been co-funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

REFERENCES

[1] Dobre, C, and C. Stratan. 2004. "MONARC Simulation Framework", in *Proc. of the 3rd Edition of RoEduNet International Conference*, Timisoara, Romania.

[2] Dobre, C, and V. Cristea. 2007. "A Simulation Model for Large Scale Distributed Systems", in *Proc. of the 4th International Conference on Innovations in Information Technology*, Dubai, United Arab Emirates.

[3] LHC Experiments, Official webpage, last accesed May 23, 2011, from http://public.web.cern.ch/public/en/lhc/LHCExperiments-en.html.

[4] Legrand, I.C., C. Dobre, R. Voicu, C. Stratan, C. Cirstoiu, L. Musat. 2005. "A Simulation Study for T0/T1 Data Replication and Production Activities", in *Proc. of the 15th International Conference on Control Systems and Computer Science (CSCS15)*, Ed. Politehnica Press, Bucharest, Romania, pp. 131-135.

[5] Bonacorsi, D., and the CMS Collaboration. 2007. "The CMS Computing Model", In *Proc. of the 10th Topical Seminar on Innovative Particle and Radiation Detectors*, Nuclear Physics B - Proceedings Supplements, Vol. 172, pp. 53-56.

[6] Casanova, H., A. Legrand, and M. Quinson. 2008. "SimGrid: a Generic Framework for Large-Scale Distributed Experimentations", in *Proc. of the 10th IEEE International Conference on Computer Modelling and Simulation* (UKSIM/EUROSIM'08), Cambridge, England.

[7] Buyya, R., and M. Murshed. 2002. "GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing", The Journal of Concurrency and Computation: Practice and Experience (CCPE), Vol. 14, pp. 1175–1220.

[8] Cameron, D.G., R. Carvajal-Schiaffino, A.P. Millar, C. Nicholson, K. Stockinger, and F. Zini. 2003. "Evaluating Scheduling and Replica Optimisation Strategies in OptorSim", in *Proc. of the 4th international Workshop on Grid Computing,* IEEE Computer Society, Washington, DC, pp. 52.