EPJ Web of Conferences **108**, 02033 (2016)
DOI: 10.1051/epjconf/201610802033

The JINR Tier1 Site Simulation for Research and Development Purposes

V. Korenkov[1,2,a], A. Nechaevskiy[1,b], G. Ososkov[1,c], D. Pryahina[1,d], V. Trofimov[1,e], A. Uzhinskiy[1,f], and N. Voytishin[1,g]

[1] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, Dubna, Russia*
[2] *Plekhanov Russian University of Economics, Moscow, Russia*

Abstract. Distributed complex computing systems for data storage and processing are in common use in the majority of modern scientific centers. The design of such systems is usually based on recommendations obtained via a preliminary simulated model used and executed only once. However big experiments last for years and decades, and the development of their computing system is going on, not only quantitatively but also qualitatively. Even with the substantial efforts invested in the design phase to understand the systems configuration, it would be hard enough to develop a system without additional research of its future evolution. The developers and operators face the problem of the system behaviour predicting after the planned modifications.
A system for grid and cloud services simulation is developed at LIT (JINR, Dubna). This simulation system is focused on improving the efficiency of the grid/cloud structures development by using the work quality indicators of some real system. The development of such kind of software is very important for making a new grid/cloud infrastructure for such big scientific experiments like the JINR Tier1 site for WLCG. The simulation of some processes of the Tier1 site is considered as an example of our application approach.

1 Introduction

The distributed complex computing systems for data storage and processing are in common use in the majority of scientific centers. For example, such systems are used in HEP (High Energy Physics) experiments where particle accelerators produce data volumes up to hundred petabytes per year. The most known experiments are LHC-CMS, LHC-Atlas and in the process of the development or design are FAIR-PANDA, BES-III and NICA-MPD. The experiments that have developed a grid-infrastructure or cloud computing for distributed data processing have some common features: huge data volumes, a long cycle of designing and construction and a long operation period. The essence of

[a] e-mail: korenkov@cv.jinr.ru
[b] e-mail: nechav@jinr.ru
[c] e-mail: ososkov@jinr.ru
[d] e-mail: pry-darya@yandex.ru
[e] e-mail: tvv@jinr.ru
[f] e-mail: zalexandr@list.ru
[g] e-mail: nikolay.voytishin@cern.ch

Article available at http://www.epj-conferences.org or http://dx.doi.org/10.1051/epjconf/201610802033

the distributed computing is that all the information from the experiment detectors comes to a huge number of data centers. The LHC grid-infrastructure represents a hierarchical structure with computer centers of Tier 0/1/2 levels [1].

The large-scale grid structure design means both the involvement of specialists possessing unique skills and the application of simulation tools. In view of the complexity of the connections, the variety of components and the large scales of the hierarchical infrastructure a simulation modelling is needed.

The urgency of the subject is caused by the fact that, in the future, the model will serve as a basis for recommendations and a requirements list for improving and developing the computer infrastructure and consideration of various variants of organizing data storage of the experiments.

2 Simulation description

The authors intend to improve the efficiency of the grid/cloud system development by using the work quality indicators of some real system to design and predict its evolution. To carry out this idea the simulation program is combined with a real monitoring system of a corresponding grid/cloud service through a special database, where the monitoring information is processed to provide necessary simulation parameters.

2.1 Analytical or imitative simulations

There are several approaches to the analytical simulation of the grid and cloud systems that can be grouped in two types: 1) The system is considered as a multi-channel queuing system with a state controlled by Markov processes under restrictions on input stream distributions and priority discipline; 2) The system is considered as a dynamic stochastic network, described by a system of equations that allow one to consider both routing and resource allocations in a network. Equilibrium and nonequilibrium network states are studied.

Both approaches give simulation results in the form of asymptotic distributions, and due to the limited theoretical assumptions cannot be applied to simulate a complex multi-tier architecture of the computer networks with real distributions of the input task streams, intricate multi-priority service disciplines and dynamic resource allocations. Therefore, the authors choose an imitative simulation method oriented on the knowledge of the system functioning dynamics.

2.2 Brief survey of grid/cloud simulation tools

There is a variety of software tools for grid systems and cloud simulation [2, 3]. For example, GridSim – class library designed to build the grid system model. It is based on the standard library SimJava which can be used to simulate the flow of discrete events in time.

The cloud computing centers can be defined as a type of parallel and distributed systems, which consist of a set of interconnected and virtual machines that are provided dynamically as one or more combined computing resources based on service level agreements (SLA) through a contract between a service provider and a consumer. The CloudSim, iCanCloud, CReST and other software toolkits are used for the cloud infrastructure simulation. These software packages allow one to create models of cloud systems with specific functionalities and configurations. The generated model runs with a simulated job flow. Thus, as a result of such simulation one can obtain a needed statistical information about the most important parameters: the average time of job execution, the life cycle of virtual machines, the resource usage. However the above mentioned simulation software is focused on simulations of specific cloud level. CloudSim functionality provides the most detailed simulation of the

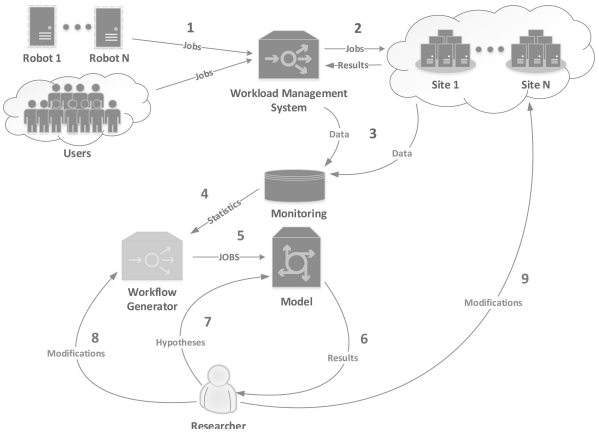


Figure 1. The simulation scheme

SaaS and IaaS levels. For the analysis of the PaaS and SaaS cloud levels iCanCloud can be used. The development of a data center with minimal electricity cost and cooling efficiency can be realized with CReST.

At the same time, the presented simulators solve quite specific tasks and do not possess a full set of functions for the grid/cloud computing and the data storage centers simulation.

Our software solution is based on the extension of GridSim classes and their combination in a program that simulates the job flow processing by grid/cloud infrastructure. Including the cloud structures into the grid allows one to reduce the solution time of the wide circle of problems and improving considerably the efficiency of the resource usage. Therefore, the methods and tools that are being developed within the project allow one to simulate the combined infrastructures including both the grid and cloud systems.

2.3 Simulation approach

The continuous development of the grid/cloud systems requires continuous adjustments of the simulation parameters. It is necessary to predict the behaviour of the system with significant changes. It is possible to use the exploitation system statistics obtained from the monitoring tools. The authors statements for the new computational system simulation are as follows: 1) hypotheses about the input data flow parameters and procedures of their processing; 2) analysis of event time distribution generated by the input data flow processing; 3) comparison of the obtained distributions with the monitoring results of the existing system.

The figure 1 shows a proposed simulation scheme. Monitoring data of real grid/cloud systems come into the database as follows: Jobs are submitted to the Workload Management System from different sources (1), then Sites are getting the jobs for execution (2), job status information enters the database (3). The Statistic data are used by Researcher to generate simulation workflow (4) then Jobs are submitted to the model (5). Researcher gets a results from the model and analyzes it (6), then he can modify the workflow parameters and verify new hypotheses (7,8). Simulation results can be used to initialize the procedure of the site configuration changes for improving its characteristics (9).

The simulation program is based on a substantially modified GridSim toolkit [5]. The events in the model are attached to the internal time as accepted in the discrete event simulation. The output is a time value of the packet processing. Understanding how the computing structure and its parts affect this time is a must. Another question to be answered by the simulation is whether are there

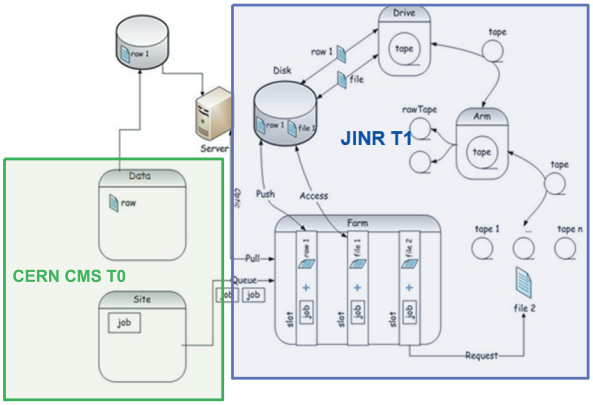


Figure 2. Tier1 – jobs and data flow scheme

any reserves of the computing system. The simulation objects are jobs, processors (slots), files, tapes, disk storages, channels, robotised libraries. The list of the events objects includes: submission jobs to the queue, retrieval jobs from the queue, occupation or release computing resources, the file transfers, extraction/mounting/ reading/writing the tapes, and others.

The last (but not least) considerable modification of the simulation program, called SyMSim (Synthesis of Monitoring and Simulation) was as follows: new classes are invented to declare data storage specific for virtual cloud clusters; input job stream is formed via data base; data exchange process is modified from a packet flow simulation into a file transfer simulation; handling of simulation results is taken out of the simulation program and is instead performed by Excel tools.

3 JINR Tier1 simulations

The years of experience with the WLCG [4] operation have shown that the only way to store the huge volumes of data produced by detectors needs robotized libraries. The problem is to simulate a data storage system with a robotized tape library, where RAW data are transferred from the disks of the great HEP experiment. A scheme of such an infrastructure is given in figure 2. The simulated structure is designed for data processing of the physical experiment, but other structures associated with the storage and updating of large amounts of digital information can also be simulated.

3.1 Model description

The model of interest here is a structure for data storage in a robotised library with thousands of tapes. The modelled object involves a number of WLCG sites interconnected by communication channels. Only one site is modelled in detail. The site consists of processing nodes, disk pools and a robotized tape storage. Other sites act as data storages and differ from the disk pools only in the properties of communication lines. The communication lines are divided into external and internal ones. The internal ones have a constant throughput. The throughput of the communications between the sites is a random variable modelled by an exponential law.

Data and tasks arrive at the site. There are three types of jobs: modeling, reconstruction and analysis. The different types of jobs share a common flow in a predetermined proportion and have different requirements to the data: 1) no input data is required for modeling but modeling generates output data; 2) no data is generated or required for analysis; 3) input data is required for reconstruction.

When a job needs data for some processing operation then if a slot and a file are available the job executes at the farm. If a file is stored in the tape library, the job reserves a slot, but is waiting for a necessary file to the disk. The tape robot moves a cartridge to the drive, the cartridge file system is mounted to the drive and the file is copied to the disk. Such a design of the data storages is assumed for the CMS Tier 1 at JINR and in the future for the NICA Tier 0/1. An illustrative example to the prospective scheme of the job flow for the CERN-CMS Tier0 – JINR-CMS Tier1 centers is given in the figure 2.

The parameters of the modelled structure are close to those of the JINR Tier1 site and are as follows: 1) Farm CPUs – 2×1200; 2) Disk pools – 5; 3) Interconnected sites – 5; 4) Pool-Farm Throughput – 900 MB; 5) Site-Site Throughput – 200 MB; 6) Drives – 8; 7) Reading speed – 160 MB/s.

In this paper the authors do not design a data storage system but demonstrate the simulator possibilities. The main parameter which affected the simulation process was average Data Acquisition (DAQ) rate. The average DAQ rate in our example is one file every 7 seconds. It coincides with the file occurrence frequency of the future NICA-SPD detector.

There are some simplifications in the model: the number of the active sites is limited; the same job flows are used for different strategies; each task demands one file only; a few tasks can use the same file; at start the files are statically distributed between the sites, disks and tapes; the files which are written on the site disk pools remain there until the experiment is over; each file has only one copy.

To determine the input parameters of the model data monitoring is required. The job flow is generated based on the distributions obtained from the statistical analysis of the data available for the JINR Tier1 site [6]. The job flow characteristics were received from the dashboard monitoring of the T1_JINR_RU site for 24 hours starting from 2015-07-09 08:00:03. A comparison of variation of the real data with the modelled ones in terms of the number of completed tasks is given in the figure 3. The agreement with the correction on shift of the model data at the size of the average consumed processor time can be seen.

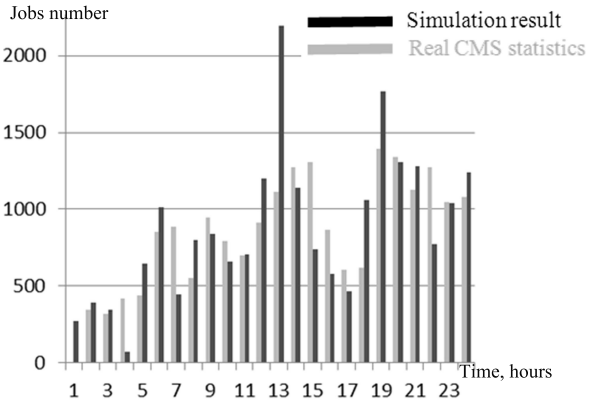


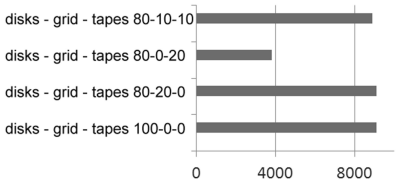
Figure 3. Real and model data on number of complete tasks for 24 hours

3.2 Simulation results

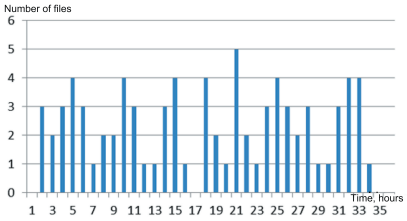
File distribution strategies have been analyzed on various infrastructure components assuming of various quality of the communication lines throughput. The number of jobs finished in a predetermined

period of the experiment defines the strategy quality. Four strategies were considered: 1) all files are on the local disks; 2) 80% of files are on the local disks and 20% are on other sites; 3) 80% of files are on the local disks and 20% are on tapes; 4) 80% of files are on the local disks, 10% are on other sites and 10% are on tapes.

The number of jobs finished inbetween 8 and 16 hours of the experiment (a) and the distribution of tapes loading (b) are shown in figure 4. Clearly, the storage on tapes with no files grouping significantly slows down the performance. On the other hand, the quality of the communication line has no much influence on the performance. The simulation has shown that the time of the job package passage without file grouping is unsatisfactory, due to the long searching time of the files located on tapes.



a) Files distribution strategies comparison



b) Distribution of tapes loading

Figure 4. Simulation results

4 Conclusion

A simplified model of the JINR Tier1 site has been developed and tested. The originality of the proposed simulation approach stems from the combination of a simulation program with a real monitoring system through a special database. The program structure is sufficiently general and flexible. It allows one to use the simulation program for the solution of design problems of data repositories which are not limited to the area of the actual physical experiments.

Acknowledgements

This work was supported by RFBR grants 14-07-00215 and 15-29-01217.

References

[1] V. A. Ilyin, V. V. Korenkov and A. A. Soldatov, Open Systems Journal **1**, 56 (2003). (In Russian). URL:http://www.osp.ru/os/2003/01/182414/

[2] V. V. Korenkov and A. V. Nechaevskiy, *System Analysis in Science and Education* **1** (2009). (Online, in Russian). URL:http://sanse.ru/download/22

[3] V. V. Korenkov, A. V. Nechaevskiy and A. N. Muravyev, *System Analysis in Science and Education* **2** (2014). (Online, in Russian). URL:http://sanse.ru/download/212

[4] The Worldwide LHC Computing Grid. URL:http://wlcg.web.cern.ch/

[5] Grid Simulation Toolkit For Resource Modelling And Application Scheduling For Parallel And Distributed Computing. URL:http://www.cloudbus.org/gridsim/

[6] CMS dashboard site. URL:http://dashb-cms-jobsmry.cern.ch/