Simulation Experience of NICA Data Storage System

V.V. Korenkov, A.V. Nechaevskiy, V.V. Trofimov

e-mail: nechav@mail.ru, Laboratory of Information Technologies, JINR, Dubna

Introduction

In various spheres of activity there are a lot of various large scale computing systems for data processing. The systems processing big data volumes are of particular interest.

The NICA accelerator complex is being created in the Joint Institute for nuclear research. Complex NICA comprises a NICA accelerator of heavy ions and a MPD installation (Multi Purpose Detector) that integrates detectors for studying f hot and dense strongly interacting QCD matter and search for possible manifestation of signs of the mixed phase and critical endpoint in heavy ion collisions.

MPD will produce a data with a stream intensity of tens of petabytes per year. The expected intensity of the data streams is so enormous that data arrays are characterized as very big ones. In order to store and process such data flows, the distributed grid-systems are used.

To optimize the structure of the future complex for data storage and processing, one should determine its key parameters and structure as well as check up the suggested technical proposals by way of simulation. For these purposes, an imitation model of the grid-site has been designed using the GridSim simulation package. This model allows one to estimate the resources required for offline data processing.

Data processing scheme of the NICA accelerator complex

The process of data receiving and processing looks as follows. The data going from processors of the MPD sub-detectors, are accumulated on-line by specially intended for event acquisition computer farm. After forming the event in off-line mode, through a dedicated 10 Gbps optic fiber communication line, the data are recorded on a disk storage. After a high level trigger, the selected events are recorded in RAW-files (the record speed is one file per minute of data gathering) and then a full reconstruction of events occurs. A predicted annual flow of processed events is approximately $19{\cdot}10^9$. Taking data transfer rate from the detectors equal to 4,7 Gbps, the total amount of initial data can be estimated in 30 PB annually, or 8,4 PB after compression. These estimations are grounded on the NICA CDR peculiarities and similar estimates executed for ALICE experiment [1].

The following variants can be taken as a system of physical information processing in the NICA experiment: a grid-infrastructure, a cloud computing or a hybrid architecture.

The experiments which apply a grid-infrastructure or cloud computing for data processing, have some common features: big data flows, a long cycle of designing and construction, a long operation period. The grid-infrastructure represents a hierarchical structure with computer centers of Tier 0/1/2 level. Functional distinctions of the levels of the hierarchical model are shown in Table 1.

Table 1: Levels of the hierarchical model and their functions [2]

Tier-0	Primary event reconstruction, calibration, storage of copies of full databases
Tier-1	Full event reconstruction, storage of actual databases on events, creation and storage of sets of analyzed events, simulation, analysis
Tier-2	Replication and storage of sets of analyzed events, simulation, analysis

Off-line level of physical data processing from the NICA accelerator complex can be simulated as a Tier-1 level site of the grid-architecture. When creating the model, one supposes that dCache will become a basis for constructing the data storage system. The model of processing the data stream of the Tier-1 site is based on the following data processing algorithm [3]:

1. Data appear with a predetermined frequency and are recorded on local disks of computers. The disk is cleared after transferring the data to the second level.
2. Data are transferred automatically to the second level over channels. As carriers of second level, pools of the dCache system are used which are considered in the model as a unified memory. While data processing it is supposed that at first the data come in the disk pool of the storage system and then under a local protocol they are transferred to processing nodes (WNs). No direct mount the file system on the working nodes is used.

3. For a long-term data storage a tape robot is used. The file copies are automatically created on tapes, then the files are deleted from the disk pools.

A distinctive feature of dCache configuration is the presence of not less than two storage levels, namely, hard disks and tapes. The tape storage means the automated libraries equipped with a robotized load mechanism and racks for cartridges (tapes). The capacity of such a library Q (quantity of tapes) can be determined by elementary calculations, the initial data for which will be the productivity of installation p (Gbps), time of its operation T (c) and storage capacity c (Gbps):

$$Q= p{\cdot}T/c.$$

Other questions related to creating the grid-site demand a more careful analysis and a choice of a comprehensible variant. Therefore, the developers face the following questions:

- Determination of a necessary quantity of drives (readwrite devices);
- Ways of grouping the files on tapes;
- A files recording policy.

Large-scale grid-structures design means both involvement of specialists possessing unique skills and application of simulation tools. In view of the complexity of connections, a variety of components and large scales, a simulation modeling is applied.

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

Simulation of the offline physical data processing in the NICA experiment

The functional specifications to the computer complex are not the same for different fields of science, experiments and user groups due to some distinctions of the computing models, needs for resources, specificity of tasks, software specialization, etc. That is why the software for simulation is designed for the needs of a particular research field. However, they have common features, and the tools created in frames of one project can be applied to a wide class of projects constructed on the grid principles. Tools and methodology for such modeling exist in plenty: GridSim, OptorSim, Monarc, ChicSim, SimGrid, MicroGrid [4]. Besides simulation languages, there is a number of projects which analyze the strategy of jobs distribution within grid-structures in order to optimize the parameters of the computing process, for example, ALEA [5]. Some tools are developed for the job flow processing simulation in cloud systems, for example, CloudSim [6].

Quite a number of systems for the simulation model development were carefully analyzed, and the GridSim platform [7] has been chosen. A number of key features of GridSim should be noted which demanded completion due to their discordance with the model requirements:

- only user can create files;
- all objects of simulation are integrated into the network by means of data links;
- user can copy (create) only one file simultaneously.

Solving these questions required expansion of existing classes and adding new objects. For instance, the following objects were added into the system:

- Drive – drive of the tape recorder;
- Arm – a hand of the robot;
- ReelArchive – cartridges archive;
- Reel – cartridge.

A set of these classes allows one to simulate all the processes occurring with a file copy on tapes: load and unload of tape by manipulator, assembling on drive, search for a file on the tape and its reading/writing.

The problem of the network infrastructure simulation in the GridSim library has been solved with the help of classes Router, Link, NetPacket and others. This suite of tools allows one to simulate the package transmission over the network. The user gets the opportunity to build-in his package schedulers into the initial model. Such an approach provides high simulation accuracy. Its disadvantage with respect to the Tier-1 problem of modeling is redundancy - no issues of routing, packages collisions, influences of the background load of channels in the given model are considered. Therefore, the level of details up to the package looks redundant. In the analyzed case, only change of load on separate components of the network is of interest. Taking this into account, a Stange class was included which allows simulations of data transfer operations. It is a reading/writing of a part of file or the whole file of experimental data.

Simulation results are accessible to the user in the form of tables and diagrams. For this purpose, classes of a log generator and imaging of results are used:

- Info – description of the computing structure and job flow;
- Reporter – log generator;
- Log parser;
- Object of imaging the results and others.

Practical usage of the Model

The usage of the mentioned classes can be illustrated by an example of modeling a process of data processing with simultaneous writing on tapes. The task of the designer is to determine a necessary quantity of the library drives. Here two questions need answers: how many library drives are required to write the whole flow of raw data (RAW) from the experiment's detectors, and to what extent the process of data processing (the job flow from users) will interfere writing if the processing demands loading files from tapes on disks.

We admit that at ones disposal there is a library, the quantity of drives in the library is fixed and is equal to five. It is essentially less than necessary, but enough to illustrate opportunities of the model. When for servicing jobs coming to the site and writing RAW data the same pools (drives) are used, the process starts behaving chaotically, repeatedly assembling and disassembling tapes for writing even at even insignificant loads. To avoid this situation, in the considered model the pools of tapes are divided into accepting data (RAW) and serving a job flow ones (DLT). The question arises how to distribute the drives between two pools at the fixed parameters of the job flow? We suppose that the files are requested in a random way.

The simulated system is a two-level one. At the first level there is a disk array, at the second one - a tape storage. In the existing model, the rate of writing and reading from the disk array does not depend upon load. The parameters of the drives and the robot correspond to the parameters of the devices planned to putting into operation (Table 2). The quantity of drives in the robot is fixed, and there is only one "hand" loading files into the drive.

Table 2: Parameters for the tape library simulation

Parameter	Value
Time of assembling / disassembling, s	22
Search rate, s	300
Reading/ writing rate, s	120
Rewind speed, s	1000
Time of load/unload of cartridge in the drive, s	100
File size, Mb	6000

Simulation results are presented in Table 3. The following characteristics were analyzed with the help of the model:

- Execution time – astronomical time of the job flow performance which will decrease with increasing the quantity of drives;

Table 3: Simulation results

Experiment	Quantity of drives RAW	Quantity of drives DLT	Execution time, s	Queue length
1	1	4	28 959	13
2	2	3	28 703	1
3	3	2	28 814	1
4	4	1	59 275	1

- Queue length – maximal length of the queue for writing the RAW data on the tape.

The simulation has shown that at the predetermined rate of data acquisition, not less than two drives should be allocated for writing. On the other hand, for the job flow processing not less than two drives should be allocated for reading the accumulated information. If one takes a minimal astronomical time of the job flow performance as a criterion of optimality, we can take the drive distribution by variant 2 as optimal.

This example illustrates one of the variants of using the program. Such investigations can be performed with the help of analytical models of the waiting theory. However, addition of some elementary conditions of grouping jobs and files considerably complicates analytical models, whereas for the simulation model the changes are reduced to several lines of the program code.

Conclusions

The created simulation system allows one to perform various experiments with an object under study with no physical realization. This provides a way for prediction and prevention of a large number of unexpected situations on-stream which could lead to unjustified expenses, loss of data and to damaging expensive equipment. In the process of simulation one can select a minimally necessary equipment which meets the requirements of data transfer, data processing and data storage, as well as estimate a necessary margin of the equipment performance that provides a possible growth of the production needs, choose several variants of the equipment in view of the current needs and the development prospects in the future, test the system operation thus revealing its bottlenecks, etc.

Application of the simulation system will allow one to determine the parameters of the data processing system of the NICA accelerator complex at a stage of engineering design.

The further development of the system assumes some additions with the purpose of creating a model of a Tier-1 level grid-site using two and three dCache levels. For simulation one supposes applying an original algorithm of the dCache pool assignment and original data on streams. Also it is necessary to perform full-scale tests of the model with the purpose of revealing errors and creating a simulation scripts base.

A particular importance of the developed simulation system is connected to the creation at JINR of an automated system of data processing and storage (ASDP) of the Tier-1 level for the CMS experiment on the Large Hadron Collider intended for work within the structure of the global grid-system for data processing (WLCG). The ASPD is aimed at performing a full cycle of processing physical information in the course of the experiment, provision of work on the physical processes simulation, a protected storage and data taking or transfer to the other WLCG centers. The dCache serves as a basis of the data storage in ASPD. Clearly, during quite a durational (10 years and more) functioning of the center, there will be a need for operative scaling of the storage system and raising the efficiency of using the tape robot in the dCache system with no interrupting the operation of the whole complex. In this process, a preliminary simulation of functioning the storage system becomes a necessary tool.

The working data can be recommended for using at designing a grid-system for acquisition, transfer, processing and storage of the data from mega-installations or other similar installations generating enormous data volumes.

References

[1] P. Cortese et al : “ALICE Technical Design Reort of the Computing.” // CERN/LHCC 2005-018, ALICE TDR 12, 2005.

[2] V.A. Ilin, V.V. Korenkov, A.A. Soldatov “Russian segment of the LCG global infrastructure” // Open Systems Journal, ISSN:1028-7493, Iissue 1, 2003, P. 56-60.

[3] Korenkov V.V. et al. Development of a simulation model of experimental data acquisition and processing at the NICA accelerator complex// Informatics and its Applications, v.7, issue 3, pp.130-137, 2013 (in Russian).

[4] A.V. Nechaevskiy, V.V. Korenkov “DataGrid simulation packages” // System Analysis in Science and Education (Online), ISSN: 2071-9612, Issue 1, 2009.

[5] ALEA- GridSim based Grid Scheduling Simulator web-portal: http://www.fi.muni.cz/ xklusac/alea/

[6] CloudSim web-portal: http://www.cloudbus.org/cloudsim/

[7] GridSim web-portal: http://www.gridbus.org/gridsim/

[8] Korenkov V.V. et al. Simulation of the Grid-system for off-line data processing for the NICA experiment // Proceedings of the 5-th International Conference ”Distributed Computing and Grid-technologies in Science and Education. Dubna, 2012.- ISBN 978-5-9530-0345-2. pp. 343-348 (in Russian).

[9] Korenkov V.V. et al. A model of the off-line data processing system of the NICA experiment // System Analysis in Science and Education, ISSN: 2071-9612, Inter. Univers. Dubna, 4, 2012.