Simulation approach for improving the computing network topology and performance of the China IHEP Data Center

Andrey Nechaevskiy[1*], *Gennady* Ososkov[1], *Darya Pryahina*[1], *Vladimir* Trofimov[1], *Weidong Li* [2]

[1]Joint Institute for Nuclear Researches, Laboratory of Information Technologies, 141980 Dubna, Russia
[2]Computing Center, Institute of High Energy Physics Chinese Academy of Sciences, Beijing, China

Abstract. The paper describes the project intended to improve the computing network topology and performance of the China IHEP Data Center taking into account growing numbers of hosts, experiments and computing resources. The analysis of the computing performance of the IHEP Data Center in order to optimize its distributed data processing system is a really hard problem due to the great scale and complexity of shared computing and storage resources between various HEP experiments. In order to fulfil the requirements, we adopt the simulation program SyMSim that was developed at the Laboratory of Information Technologies of the Joint Institute for Nuclear Research. 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. SyMSim facilitates making a decision regarding required equipment and resources. The simulation uses input parameters from the data base of the IHEP computing infrastructure, besides we use some data of the BESIII experiments to indicate workflow and data flow parameters for simulation three different cases of organizing IHEP computing infrastructure. The first simulation results show that the proposed approach allows us to make an optimal choice of the network topology improving its performance and saving resources.

1 Introduction

The Chinese Academy of Sciences Research Institute of High Energy Physics (IHEP) [1] is China's biggest laboratory for the study of particle physics. Dynamically developing high energy physics (HEP) experiments, such as Beijing Spectrometer (BESIII) Experiment [2], are expecting to deal with Exabyte data scale and need corresponding means of distributed computing. The development of sophisticate grid-cloud systems intended to store, distribute, and process super-big volumes of experimental data inevitably demands a substantial study of their optimality by detailed simulation of these systems.

* Corresponding author: nechav@jinr.ru

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

Simulation program SyMSim (Synthesis of Monitoring and Simulation) [3] was developed at Laboratory of Information Technology (LIT) of the Joint Institute for Nuclear Research (JINR) and then modified for the IHEP simulation.

2 Basic concepts of simulations

The simulation goal of a modern computer center is to satisfy some optimality criterion which minimizes the equipment cost under unconditional fulfilment of SLA (Service Level Agreement). The best way to evaluate dynamically the system functioning quality is using its monitoring tools. The basic concepts underlying the simulation program SyMSim suppose that this is combined with a real monitoring system of the grid-cloud service through a special database (DB) and includes the following components:

- a core – its stable main part independent on simulated object;
- a declarative module for input of model parameters defining a concrete distributed computing center,
- its setup and parameters obtained from monitoring information, as dataflow, job stream, etc.

DB intention is just to realize this declarative module work and provide means for storing and output simulation results. A web-portal is also needed to communicate with DB, assign concrete simulation parameters and output results stored in DB.

3 The simulation experience of China IHEP Data Center

As possible structure cases of the China IHEP Data Center to be compared by results of their simulation in our simulation experiments we use its simplified schema (case 1) with two possible extension (cases 1-2) depicted in Figure 1.

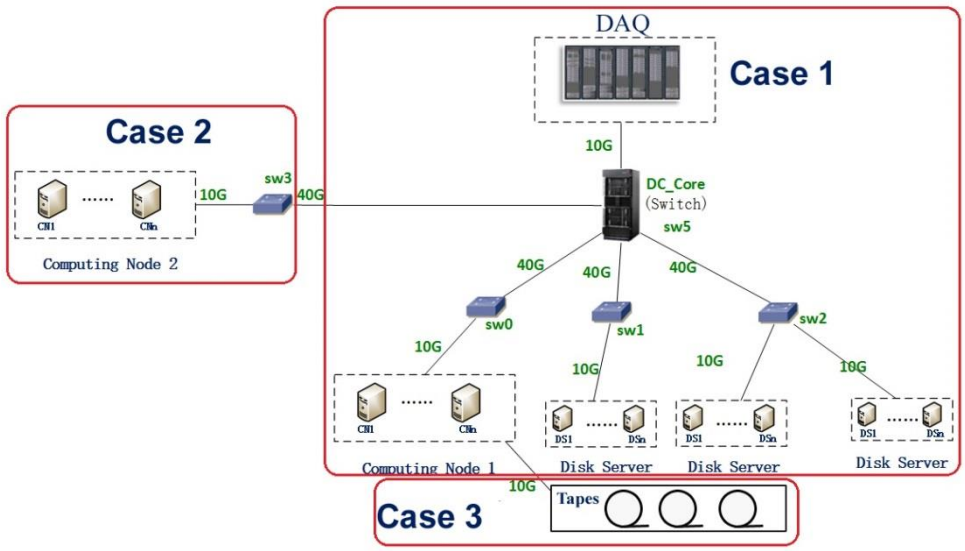

Figure 1. Simplified schema of 3 possible cases of the China IHEP Data Center with one or two computing nodes or robotized tape library extension (cases 1-3).

For the beginning we simulate the first case, as the typical process of job flow in one of Computing Nodes with 500 CPUs. A file needed to perform one or more jobs must be available on the remote Disk Server and requires downloading to a local pool. We

submitted 10000 jobs with files from 0.1 GB to 100 GB. The time, when CPUs are busy by idling jobs, because they cannot start waiting for this file from Disk Servers, can be considered as the important characteristic of the computing system loss.

Then we simulate the second case of this centre, which includes two computing nodes with 500 CPUs.

To analyse the usage level of those two cases in our simulation we vary the intensity of the data and job flow and the load of communication equipment for one and two computing nodes. Based on comparing results of our simulations one can identify problems, confirmed the quantitative characteristics that arise in the process of data processing.

Among events occurring at the system during the simulation run one can compare for considered cases such characteristics of the computing process, as the job queue dynamics, the load of switches (Figure 2) or cases of the system loss since CPUs are busy by idling jobs (Table 1). It was decided to base the comparison of considered cases on the system losses due to CPU occupations by idling jobs waiting for a file.

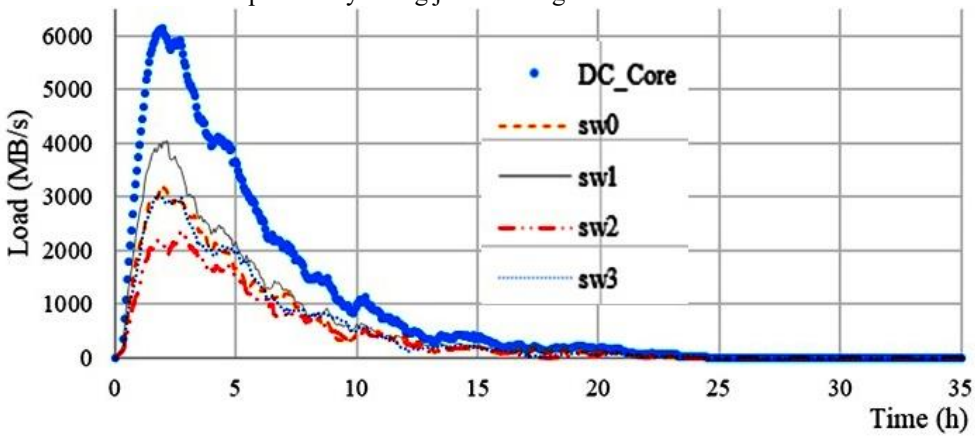

Figure 2. Load of switches.

Table 1. Comparison of time intervals between job submitting and execution start for cases 1 and 2.

	Average delay time (min)	**Number of jobs w/o delay**	**Number of jobs with delay less then 60 (min)**	**Number of jobs with delay over 60 (min)**	**Total waiting time (min)**
Case 1 (500 CPU)	7,2	8567	872	561	72209
Case 2 (2*500 CPU)	12,6	7598	1396	1006	126245

As simulations show, we have got quite the same time for all jobs proceeding. But in case 2 (Table 1) the total waiting time (time when the job waiting for a file) increased by 75%. So when 500 CPUs are added in Case2, it is not any effective way.

However, if we choose the different way of increasing the computing power and add not a computing node, but extra 500 cores to existing computing node, then on the node with 1000 cores the losses stay on the same level.

The last case 3 includes such important part of the China IHEP Data Center as a robotized tape library. The intention of its simulation is discussed in the next section.

4 Possible improvements of the job flow process

Thus it is shown that the program SyMSim is successfully adopted and allows to obtain a number of important quantitative characteristics of job flow and dataflow processes needed to see how to optimize the system.

In particular, simulation shows that attempts to increase the power of computer system by enlarging the number of computer nodes leads to increasing system losses due to idle processors. However, we can keep losses at the same level, if we would increase the computing power by enlarging the number of cores in one node.

There are, at the same time, technologically different solutions to speed up the workflow process and improve CPU usage:

- the use of cloud infrastructures and virtualization;
- develop a scheduler that will load the job to execution, taking into account the availability of the needed file(s);
- launch procedures of pre-load files.

The choice of the solution depends on the architect of the data processing system, which he can take based on the simulation results.

As the first step, the LIT team extends the SyMSim algorithm to the case 3 (see Figure 1) that includes such important component of the computing center as data stream from the data acquisition (DAQ) infrastructure to be stored on a robotized tape library. The case 3 simulations show the process of storing data from the DAQ system and from the tapes on the buffer disks at the same time. The aim of the simulation in this case is to choose number of drives in the robotized tape library we need to store all data. So we investigate the disk buffer loading dynamics under different conditions

Data acquisition systems receive and store event data from individual detector with a frequency corresponding to the output frequency of the trigger system. Such a trigger periodicity we are measuring by the trigger period, i.e. by the average time interval between sequential data transfers from the trigger via some buffer. We performed the simulation run of our experimental system for different trigger periods and calculated the dependence of the total numbers of transferred files from this period that is shown in Figure 3.

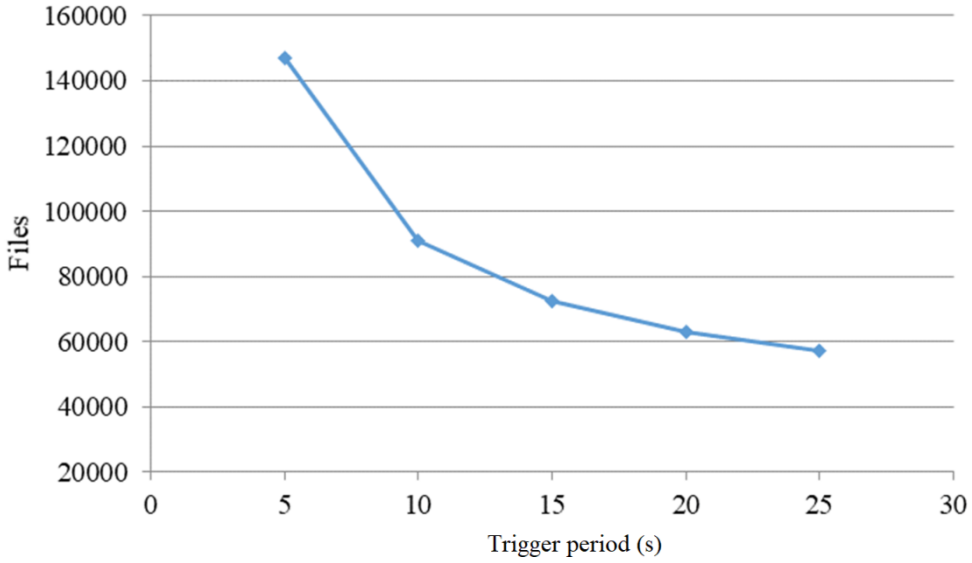

Figure 3. Dependence of total file numbers on trigger period.

Since the number of transferred files depends not only on the trigger period, but at the same time on the number of drives in the robotized tape library, we simulated this double dependency , first, for the case shown in Figure 4 when files stored to the buffer from the DAQ system without a tape copy (we cannot delete such files). The total number of files with a tape copy is shown on Figure 5. As it seen, with a sufficient number of drives (7 in our case) we can avoid any queue of files because we have time to write all the files on tapes.

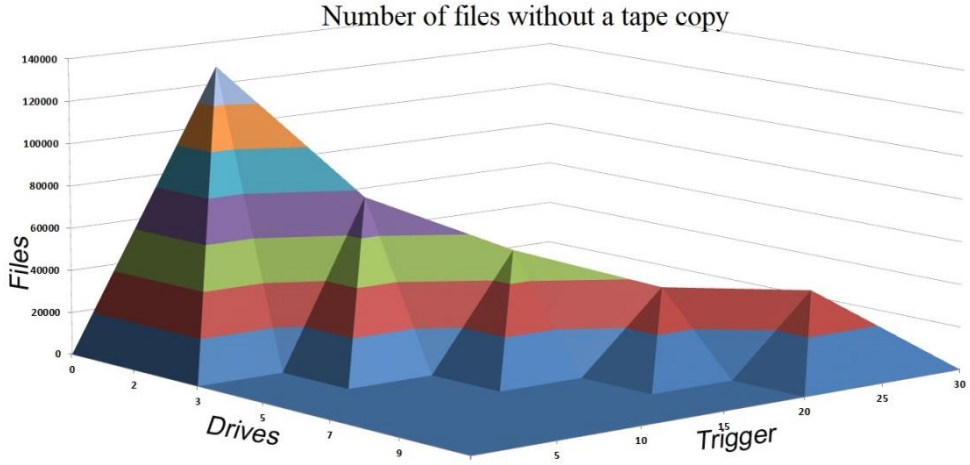

Figure 4. File number dependence on trigger period and drive number in the robotized tape library. when files stored to disk only without a tape copy.

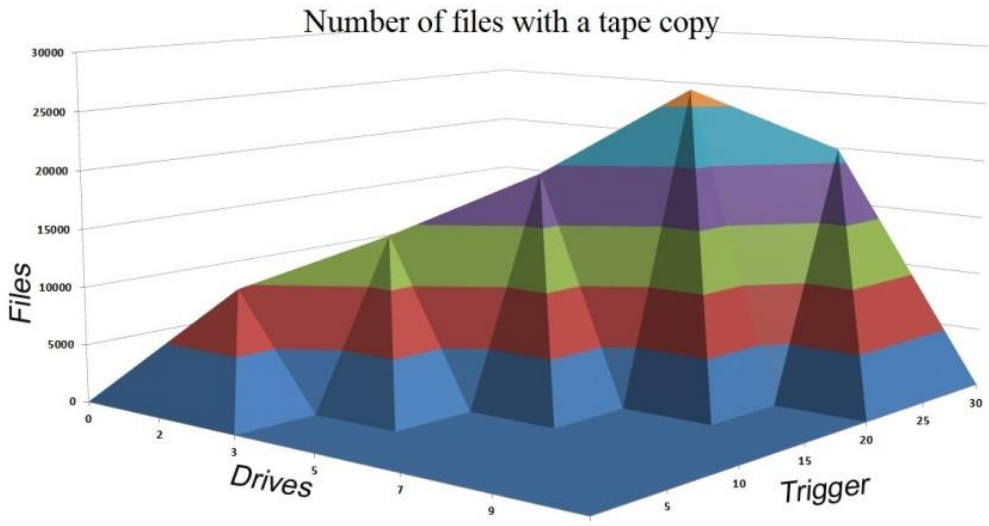

Figure 5. Dependence of the total file number from trigger speed and drives number in the robotized tape library, when files stored on disk and tapes.

5 Conclusion

Our first experience with simulating the IHEP computing is very preliminary and intended just to try to adapt an existing simulation program to IHEP specifics.

The new version of our simulation program was already installed in the China IHEP Data Center, adapted to some of its parameters and tested.

The first attempt of simulations was accomplished on the quite simplified model of the IHEP computing Center disregarding such its important parts, as DAQ infrastructure etc.

Nevertheless, since the certain success of this experience has demonstrated the applicability of the simulation program, we are going to extend the IHEP Computing Center model to be simulated gradually approaching to its present and then planned structure.

References

[1] IHEP web page (2018). *Institute of High Energy* Retrieved september 10, 2018, from: english.ihep.cas.cn

[2] BESIII web page (2018). *Beijing Spectrometer Experiment* Retrieved september 15, 2018, from: bes3.ihep.ac.cn

[3] Korenkov V. V., Nechaevskiy A. V., Ososkov G. A., Pryahina D. I., Trofomov V. V., Uzhinskiy A. V., *Simulation concept of NICA-MPD-SPD Tier0-Tier1 computing facilities* // Particles and Nuclei Letters, vol. **13**, No **5**, pp. 1074–1083 (2016).