The International Symposium on Grids and Clouds (ISGC) 2018

IHEP

Simulation approach for improving the computing network topology and performance of the China IHEP Data Center (Preliminary results)

Li Wang [1], Fazhi Qi [1], Andrey Nechaevskiy[2], Gennady Ososkov[2], Daria Pryahina[2], Vladimir Trofimov[2] , Weidong Li [1]

[1]Computing Center, Institute of High Energy Physics Chinese Academy of Sciences, Beijing, China
[2]LIT, Joint Institute of Nuclear Research, Dubna, Russia

Outline

- HEP experiments in China
- The problem:

  IHEP Data Center performance

  evaluate network performance
- Test approaches
- Netbench
  - fundamental environment
  - sample results
  - possible improvements

Large science facilities

- IHEP: The largest fundamental research center in China
- IHEP serves as the backbone of China's large science facilities
  - Operation
    - Beijing Electron Positron Collider BEPCII/BESIII
    - Yangbajing Cosmic Ray Observatory: ASg & ARGO
    - Daya Bay Neutrino Experiment
    - Hard X-ray Modulation Telescope (HXMT)
    - Accelerator-driven Sub-critical System (ADS)
  - Under construction
    - China Spallation Neutron Source (CSNS)
    - Jiangmen Neutrino Underground Observatory (JUNO)
    - Large High Altitude Air Shower Observatory (LHAASO)
    - Beijing High Energy Photon Source (HEPS)
  - Under planning
    - XTP, HERD, CEPC, …

Main Experiments at IHEP

BESIII (Beijing Spectrometer III at BEPCII)

DYB (Daya Bay Reactor Neutrino Experiment)

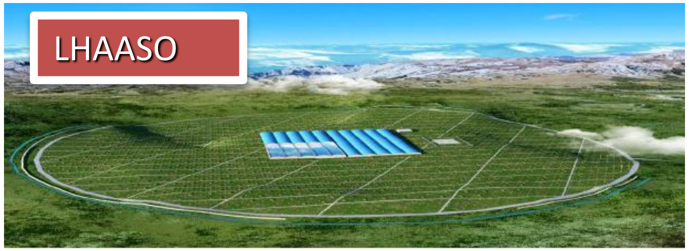
JUNO (Jiangmen Underground Neutrino Observatory)

CMS

Compact Muon Solenoid

ATLAS EXPERIMENT

Beijing High Energy Photon Source



Large High Altitude Air Shower Observatory

Hard X-Ray Moderate Telescope

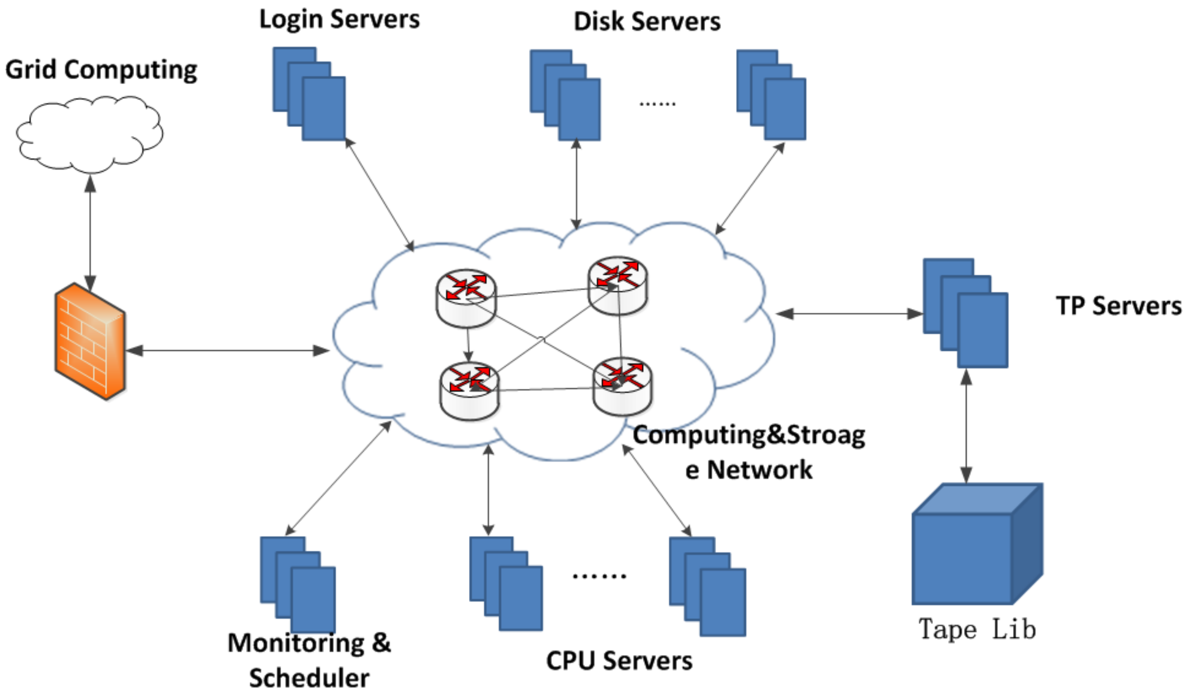
Problem1:New challenges for IHEP Data Center

Demands for the growing numbers of HEP experiments

- Hosts
- Computing resources
- Storage resources
- Network resources

It is really difficult for us to evaluate the performance of network on actual IHEP Data Center.

Problem2:Network performance is key



IHEP Computing System

Network is core

- Login servers
- Disk servers
- TP servers
- Scheduler
- CPU servers
- Etc.

How to evaluate the Network performance

Basic concepts of simulations

- The goal of basic concepts of simulations 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 simulation program is to be combined with a real monitoring system of the grid/cloud service through a special **database** (DB);
- To ensure a developer from writing the simulation program from zero on each development stage it is more feasible to accept a **twofold model structure**, which consists of
  - **a core – its stable main part** independent on simulated object and
  - 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 output of simulation results;
- **Web-portal** is needed to communicate with DB assigning concrete simulation parameters and storing results in DB.

What simulations should give us

on the design of Computing Environment

- Evaluate Computing Environment performance and reserves under various changes:
  - Different workloads
  - System configuration
  - Different scheduling algorithm
  - Hardware malfunctions
- Balance the equipment needed for data transfers and storage by minimizing cost, malfunction risk and execution time;
- Optimize resource distribution between user groups;
- Predict and prevent a number of unexpected situations
- Test the system functioning to find bottlenecks.

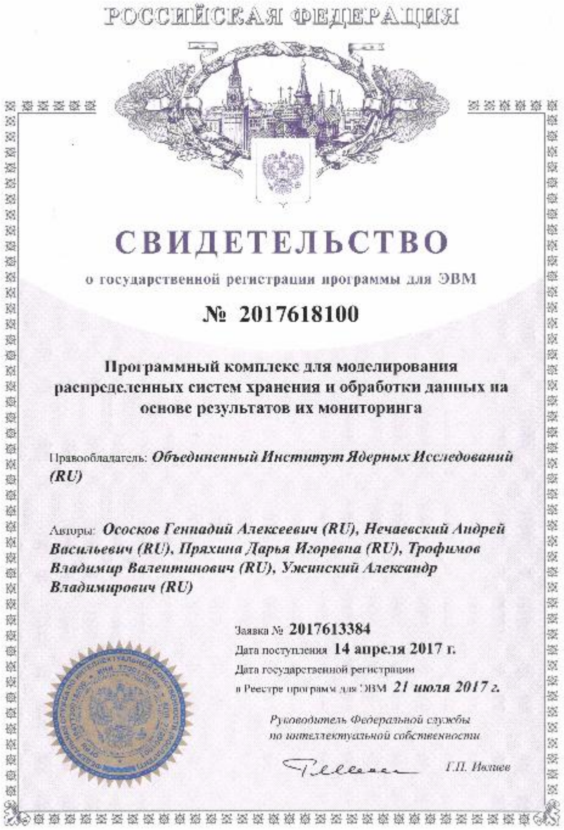
Simulation of grid-cloud systems at JINR

- The team from the Lab of Informational Technologies (LIT) of JINR Dubna has already the experience with the simulation of grid structures by the new simulation program system called **SyMSim (Synthesis of Monitoring and SIMulation)**
- The SyMSim was successfully tested for the JINR CMS Tier 1 center and then used to simulate computing facilities for T0-T1 NICA project
- In 2017 SyMSim obtained the certificate of state registration №2017618100

The SyMSim system is focused on predicting evolution of work quality indicators of the grid-cloud systems. Simulation results were already successfully used to choose an optimal grid-cloud configuration of several distributed computing centers.

The recent successful SyMSim applications were

1. Simulation of distributed data processing system for BM@N experiment of T0-T1 NICA project
2. Simulation of interprocessor interactions for MPI-applications in the cloud infrastructure



РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2017618100

Программный комплекс для моделирования распределенных систем хранения и обработки данных на основе результатов их мониторинга

Правообладатель: ***Объединенный Институт Ядерных Исследований (RU)***

Авторы: ***Ососков Геннадий Алексеевич (RU), Нечаевский Андрей Васильевич (RU), Пряхина Дарья Игоревна (RU), Трофимов Владимир Валентинович (RU), Ужинский Александр Владимирович (RU)***

Заявка № **2017613384**

Дата поступления **14 апреля 2017 г.**

Дата государственной регистрации
в Реестре программ для ЭВМ ***21 июля 2017 г.***

*Руководитель Федеральной службы
по интеллектуальной собственности*

Г.П. Ивлиев

Database design

•***Reminder:*** The simulation program is to be combined with a real monitoring system through **a special database (DB),** which intention is just to input of model parameters and output of simulation results.

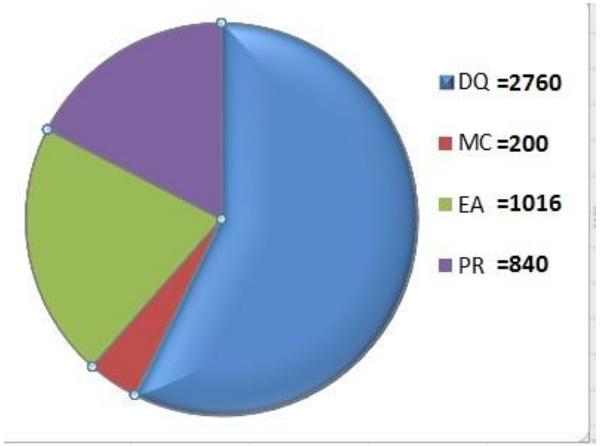
Database contains the description of the computing nodes, links between them, grid structure with its nodes, running jobs information, execution time, the monitoring results of the various subsystems of the grid and the simulation results.

Database main tables

- ❑ **Simulation_Parameters** — describes main parameters of simulation (time, input-output);
- ❑ **Configurations** — contains a description of hardware and net topology ;
- ❑ **Jobswaiting** — contains a description of a job flow and files distribution (the model of input data);
- ❑ **Results** — program results.

Four types of jobs are generated

1. Data acquisition (**DQ**) – simulated "raw" data to be stored
2. Monte-Carlo (**MC**) – do not need input data
3. Express analysis (**EA**) – jobs use recently obtained files
4. Reconstruction processing (**PR**) – jobs consume the most of resources



Example of job type distribution for JINR TO/T1

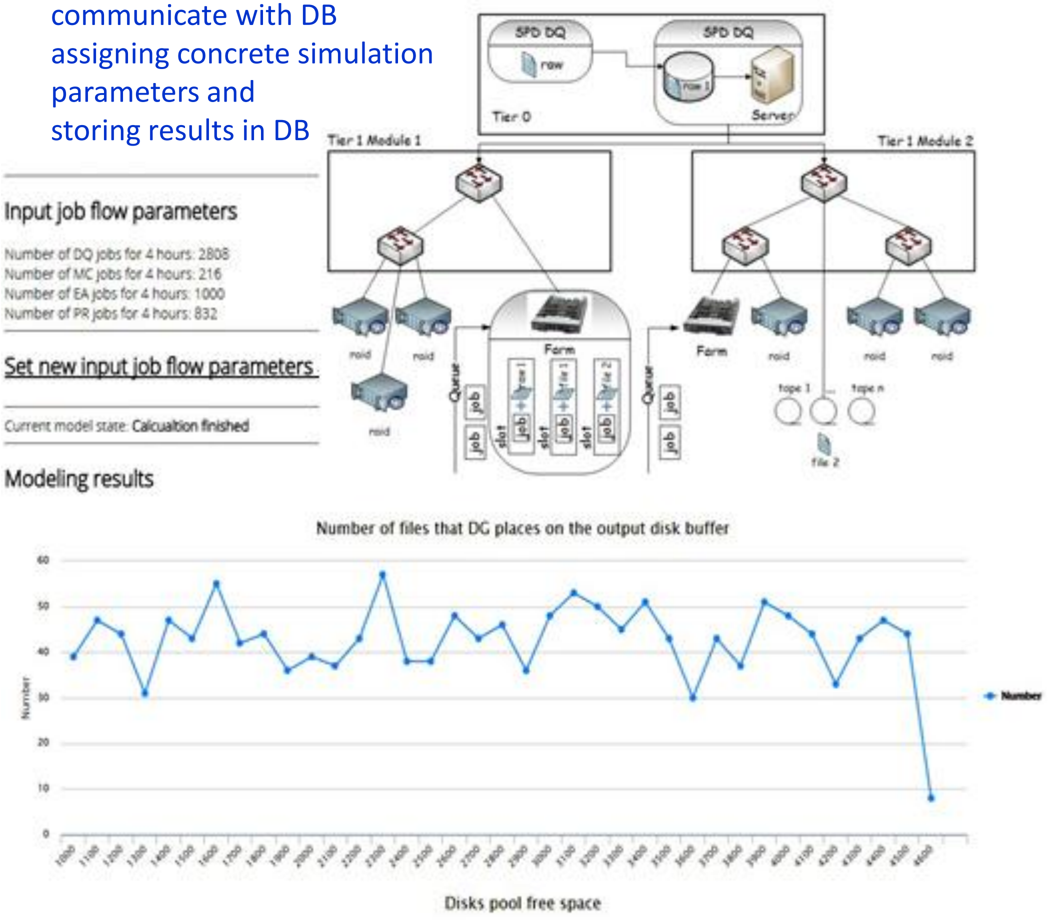
Web-portal functions

- Interaction with the database.
- Present current model structure and generated workflow description.
- Set new workflow with different parameters (number of DQ, MC, EA, PR jobs) generation.
- Simulation results representation (graphics, diagrams).

The result of the simulation program is a sequence of records in the database, which reflects all the events occurring at the system during a simulation run.

These include: the emergence of jobs, the processor allocation, start and end data processing (job start and completion), beginning and end of loading files, etc.

Reminder:
Web-portal is needed to communicate with DB assigning concrete simulation parameters and storing results in DB



Input job flow parameters

Number of DQ jobs for 4 hours: 2808
Number of MC jobs for 4 hours: 216
Number of EA jobs for 4 hours: 1000
Number of PR jobs for 4 hours: 832

Set new input job flow parameters

Current model state: Calcualtion finished

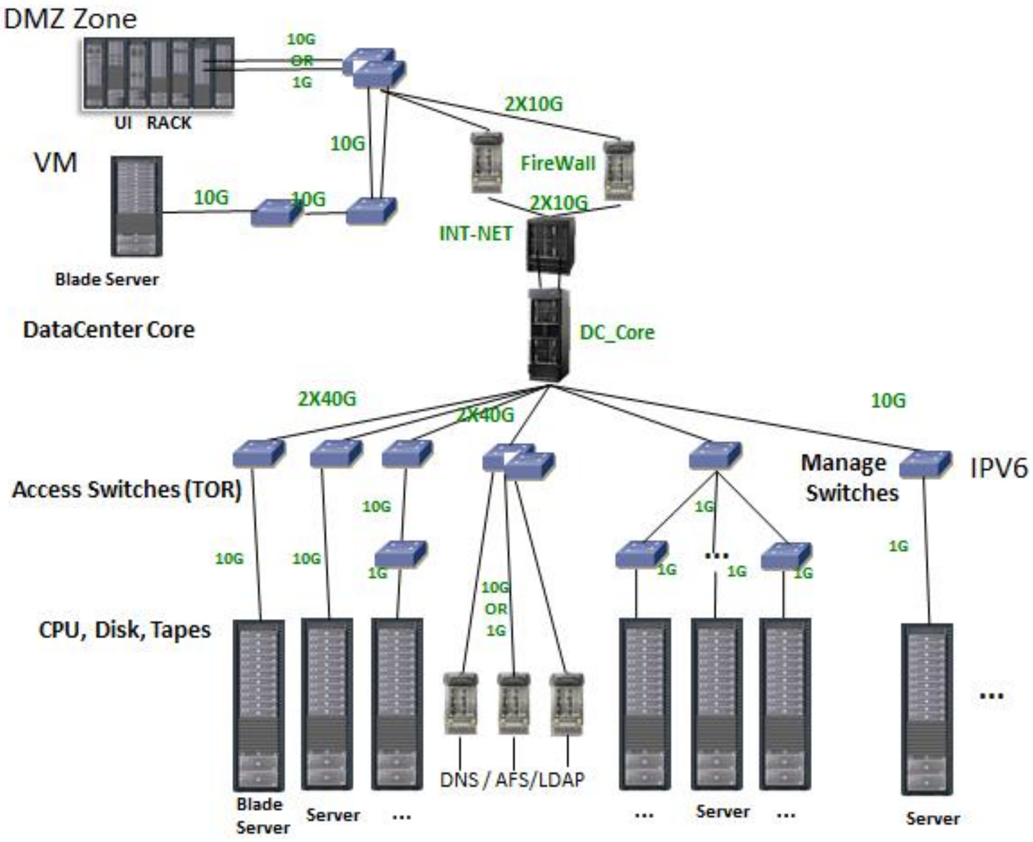
Modeling results

Snapshot of SyMSim web-portalt

Simulation algorithm is designed that at the initial time all buffers are empty, the processor is not loaded and data are not transferred. **Therefore the initial transition process must be excluded from the analysis.** It also happens when the current job flow stops.

Simulation Environment

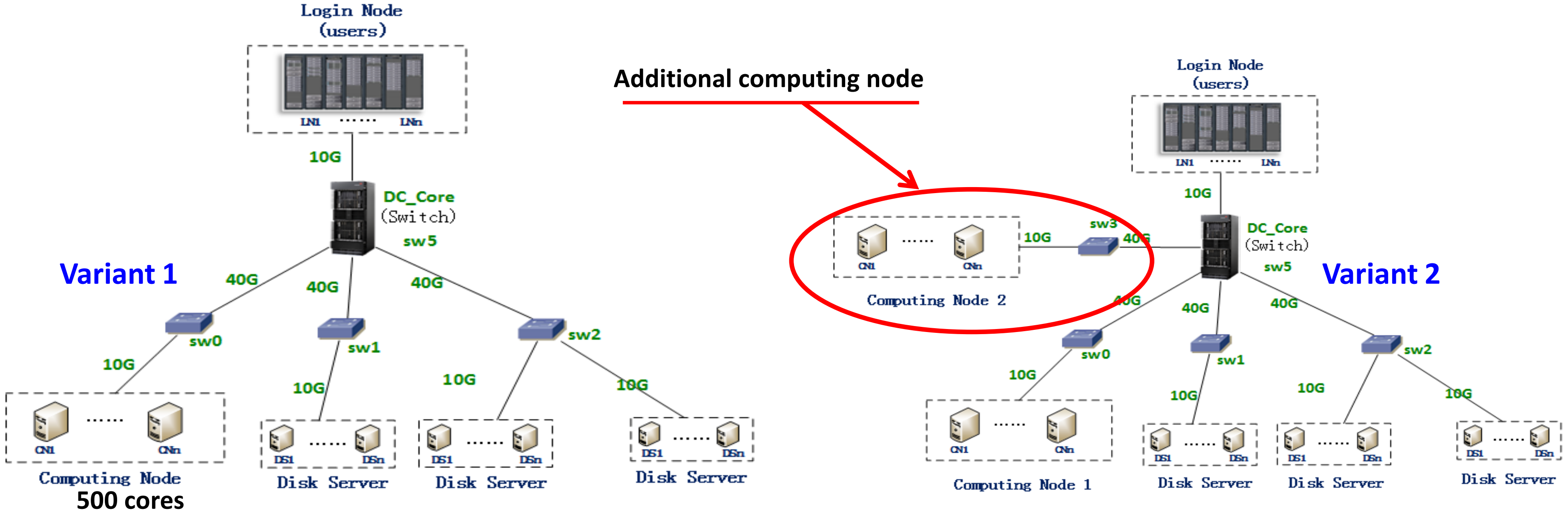
- IHEP Data Center
  - CPU Cores :6844
  - Computing Nodes:504
  - Storage
    - Hosts:24
- ❖ Job Stream
  - BESIII experiments data
- Job Type
  - Simulation
  - Analysis
  - Reconstruction
- ❖ Result
  - Preliminary results



The first simulation experience with a simplified IHEP computing scheme

After obtaining simulation results one can compare the usage level of various variants of computing node extensions by analyzing the intensity of the data and job flow and the load of communication equipment. Based on these results one can identify problems, confirmed the quantitative characteristics that arise in the process of data processing.

For the first simulation experience with IHEP computing we choose two simplified schemes



Additional computing node

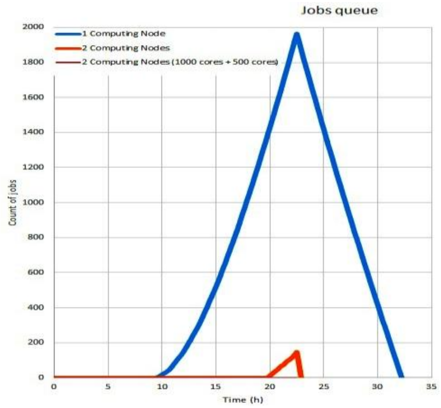
Variant 1

Variant 2

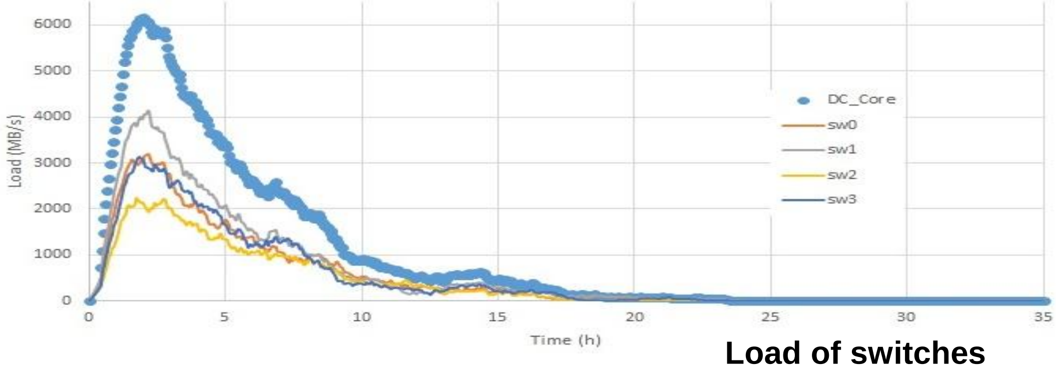
Let us consider a typical process of job flow in one of Computing Nodes with 500 PC.
A file needed to perform one or more jobs must be available on the remote Disk Server and require downloading to a local pool.
The time, when CPUs are busy by idling jobs, because they cannot start waiting for a file, can be considered as the important characteristic of the computing system loss.

Examples of computing process characteristics obtained by simulation

Among events occurring at the system during a simulation run one can compare for considered variants such characteristics of the computing process, as the job queue dynamics, the load of switches, or cases of the system loss since CPUs are busy by idling jobs.

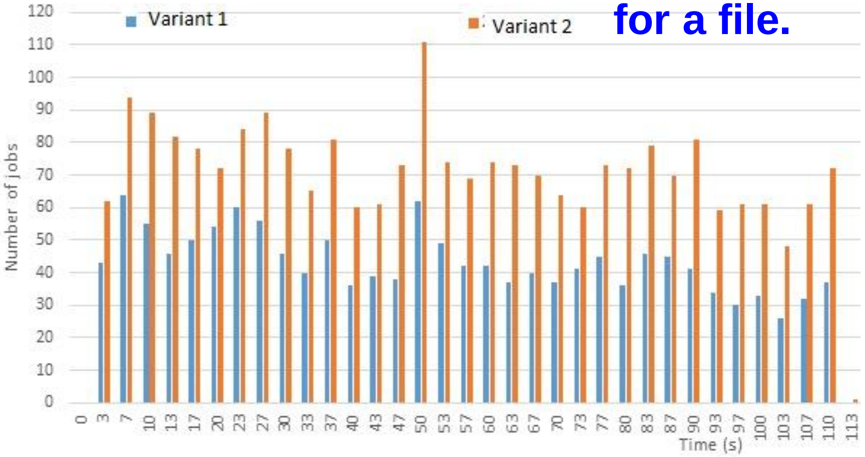


Job queue dynamics



Load of switches

We decided to base the comparison of considered variants on the system loss due to CPU occupations by idling jobs waiting for a file.



Comparison of job distributions according to the time elapsed since sending the job to the CPU to start calculations for two variants

As our simulations show, the system for variant1 looses 8%, but after adding the second computing node (Variant 2) the loss increases up to 15% of system time.

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 stays on the level 8%**.

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 jobflow and dataflow processes needed to see how to optimize the system.

In particular, simulations 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, you can keep losses at the same level, if you 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 jobflow process and improve CPU usage:

1. The use of cloud infrastructures and virtualization.
2. Develop a scheduler that will load the job to execution, taking into account the availability of the needed file(s).
3. Launch procedures of pre-load files.

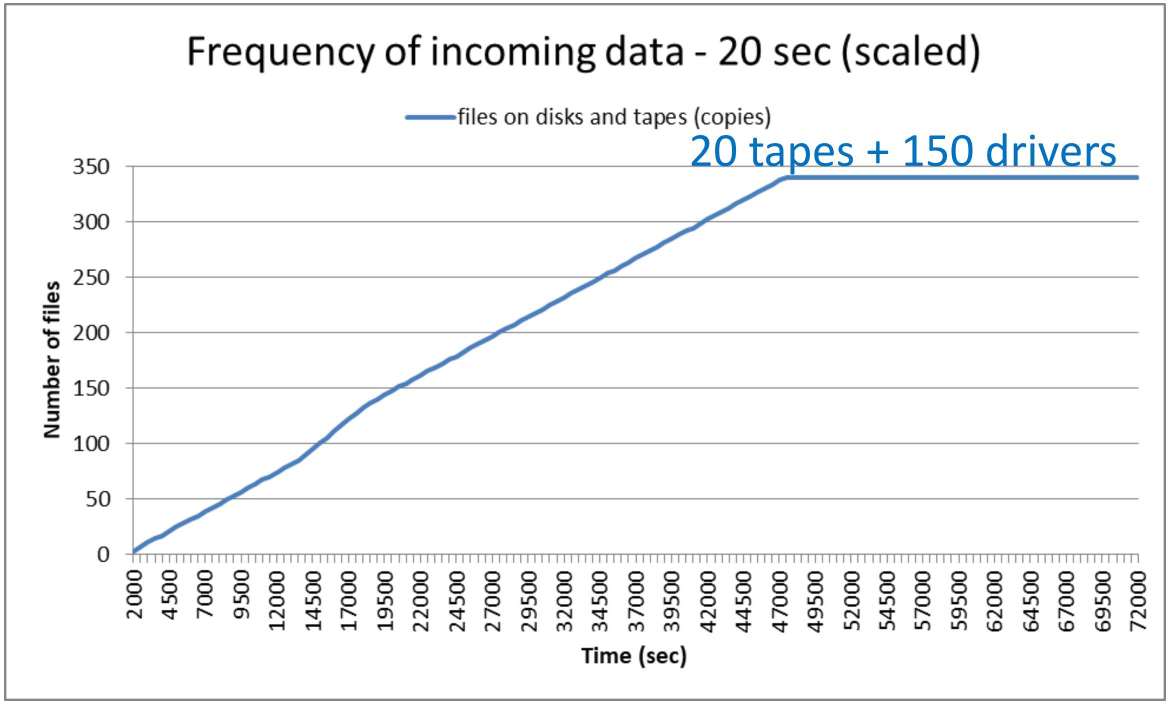
The choice of the solution depends on the architect of the data processing system, which he can take based on the simulation results.

Data stream intensity VS number of tapes

As the first step, We extended the SyMSim algorithm to include such its important parts as data stream from data acquisition infrastructure to be stored on robotized tape library.

Simulation results show:

It will be not enough only 20 tapes (2 TB storing capacity each of them) to write all incoming data flow with the frequency every 20 sec



20 tapes + 150 drivers

Conclusion and outlook

- Our first experience with simulating the IHEP computing is very preliminary and intended just to try to adapt an existing simulation program to the IHEP specifics.
- The new program version was already installed in the CC IHEP, adapted to the CC parameters and tested.
- Success of this experience has demonstrated the applicability of the simulation program, so we are going to extend the IHEP Computing Center model to be simulated gradually approaching to its present and then planned structure.

Thanks for your attentions!