Simulation of a Distributed Data Processing System for HEP Experiments

Andrey Nechaevskiy
Joint Institute for Nuclear Research

New computing challenges for Big Data scale

- The development of a HEP (High Energy Physics) computing system is a complex and difficult task due to the need of a sophisticated design under evolving user requirements.
- Potentially new physics is expected at the LHC 2nd run and the new JINR NICA megaproject
- Working at TB scale the MPD-SPD experiments will face with great challenges in distributed computing:
  - ***large increase of CPU and network resources****;*
  - ***combined grid and cloud access;***
  - ***Intelligent dynamic data placement;***
  - ***distributed parallel computing;***
  - ***renewal most of simulation and analysis software codes.***
- These problems are inherent to such the JINR projects as the running Tier 1 for CMS and the planning Tier 0/1 for NICA

Simulation of grid-cloud systems

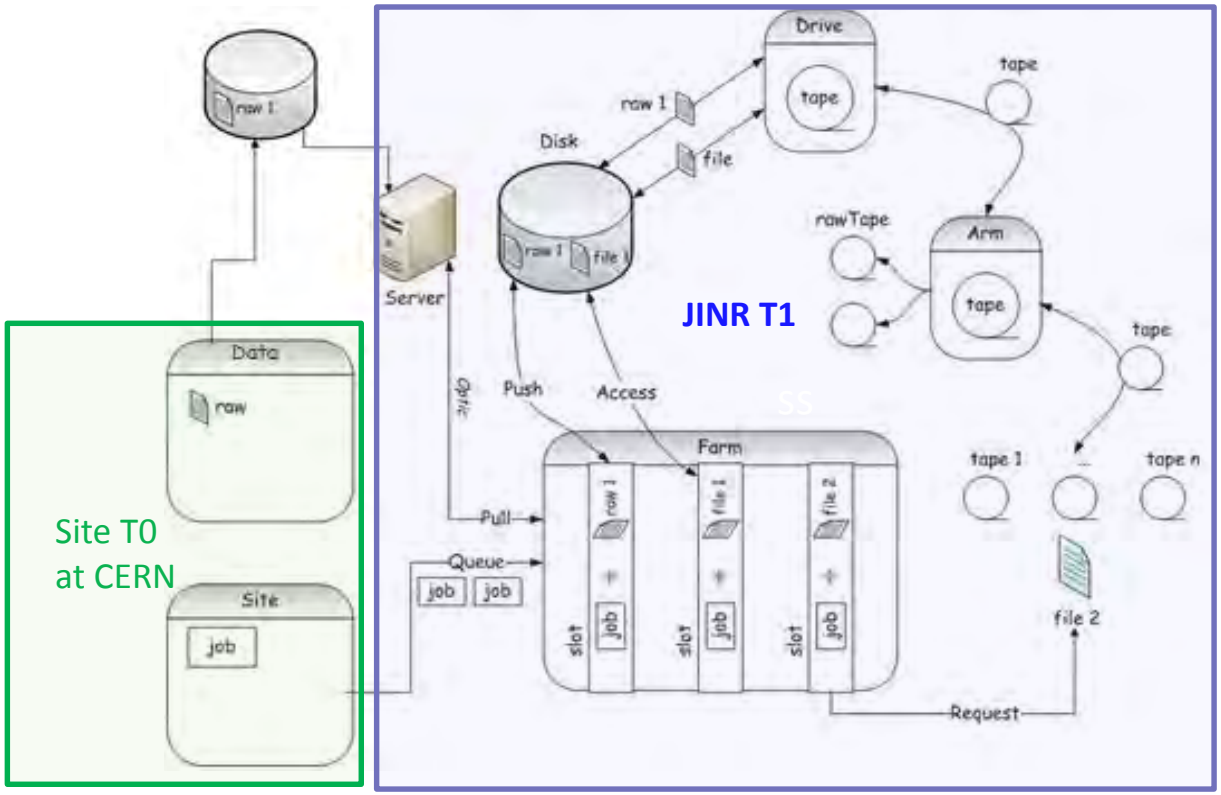
- **Substantial optimality study** is needed to avoid possible and quite expensive mistakes on design and development stages of any grid-cloud system
- The study of grid-cloud system optimality is based on the **optimality criterion** which minimizes the equipment set (cost) under unconditional fulfilment of **SLA** (Service Level Agreement)
- Such studies can be efficient when it is based on **scrupulous simulations** of
  - Job stream with knowledge of
    - Job types (simulation, analysis, reconstruction)
    - Statistical information about distribution of their arrival and execution times
  - computing resources (number of compute nodes, the architecture of a computer system, installed software, CPU consumption)
- Efficient simulation of grid-cloud systems should take into account the **functioning quality** of this system to evaluate its performance and to forecast its future.

SyMSim simulation program

- Our team has already the experience with simulation grid structures inspired by GridSim library (http://www.buyya.com/gridsim) and job scheduler ALEA (http://www.fi.muni.cz/~xklusac/alea).
- The new simulation program called **SyMSim (Synthesis of Monitoring and SIMulation)** was developed at LIT-JINR with aiming to provide necessary computing background for NICA complex.
- SyMSim allows improving the efficiency of the grid/cloud structure development accommodating the work quality indicators of real system.
- To accomplish that
  1. New classes are invented to declare the data store specific for the tape robot library;
  2. Input job stream is formed via data base;
  3. Data exchange process is modified from packet flow simulation into file transfer simulation;
  4. Software means for handling simulation results are provided.

Tier1 Dataflow simulation

The problem is to simulate a data storage system with robotized tape library, where RAW data are to be transferred from disks of a great HEP experiment. In reality such data storages are used for the CMS Tier 1 at JINR.



Scheme of the job and data flow at JINR T1

Tape robot IBM 3500

How it works on T1 site:

1. **From disk to tape:**

- If slot and file are available, job is executed at the farm;

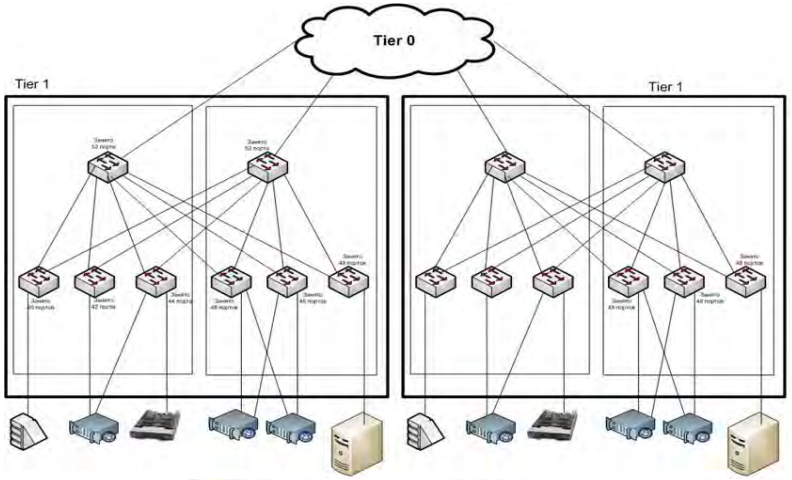
2. **From tape to disk**:

- If file stored in tape library. job reserves a slot, but is waiting for necessary file to the disk: the robot moves tape cartridge to the drive, cartridge's file system mounting to the drive, file is copied to the disk.

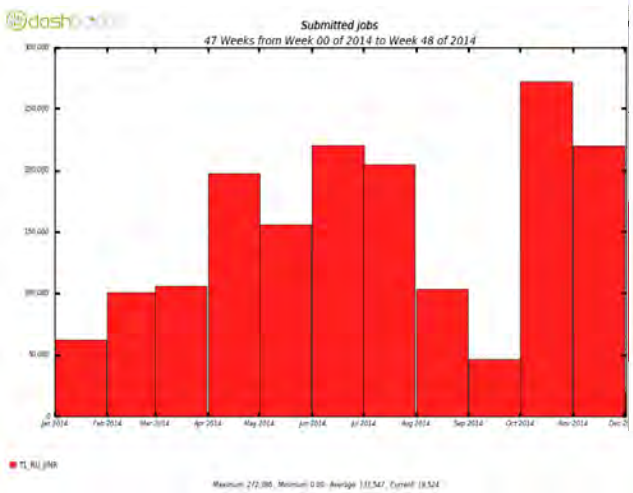
Model verification by comparing JINR Tier 1 real and simulated characteristics

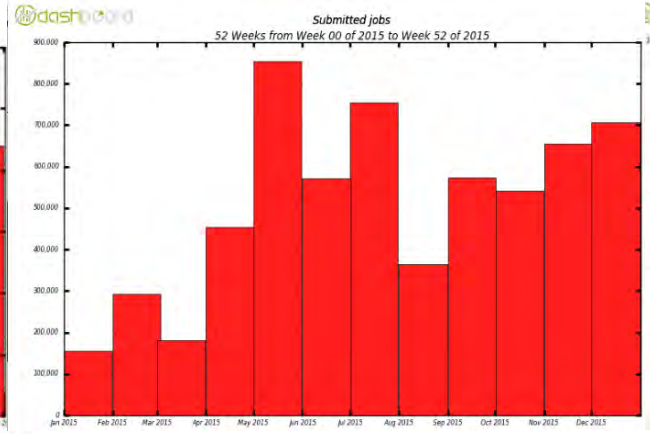
CPU - 2400
Disks - 2400 TB
Tapes - 5 PB

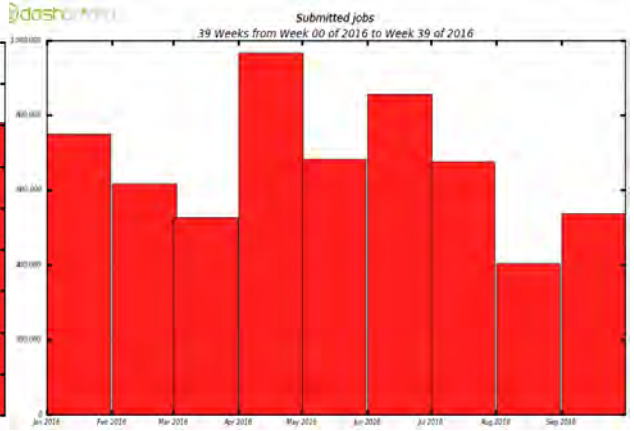
these parameters from real T1 were set to the model



Job stream for the JINR Tier1 is characterized by the following statistics.





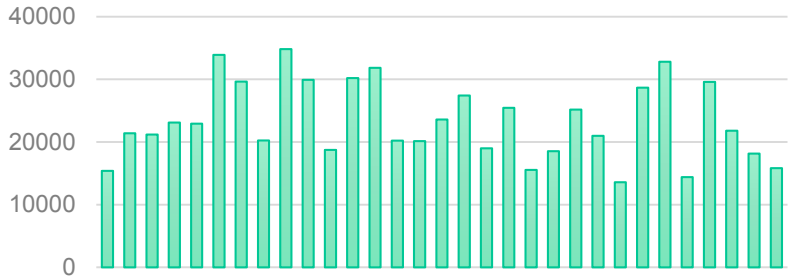


~ 2 mil. Submitted Jobs (2014) ~ 5.8 mil. Sub. Jobs (2015) ~ 6 mil. Sub. Jobs (9 month of 2016)

Statistics was taken from http://dashb-cms-job-dev.cern.ch/dashboard/

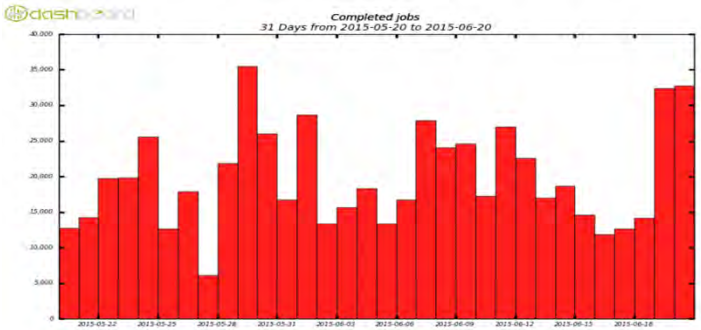
Real and Generated Workflow (CMS T1 JINR)

Completed jobs (simulated)



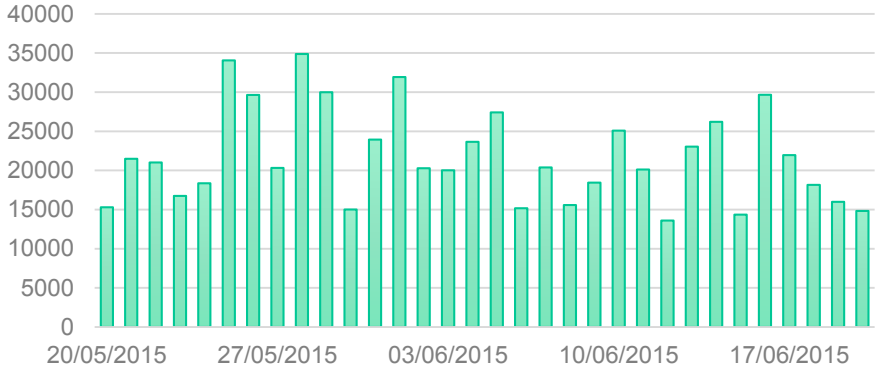
X = 24000 S = 6100

Completed jobs (real)



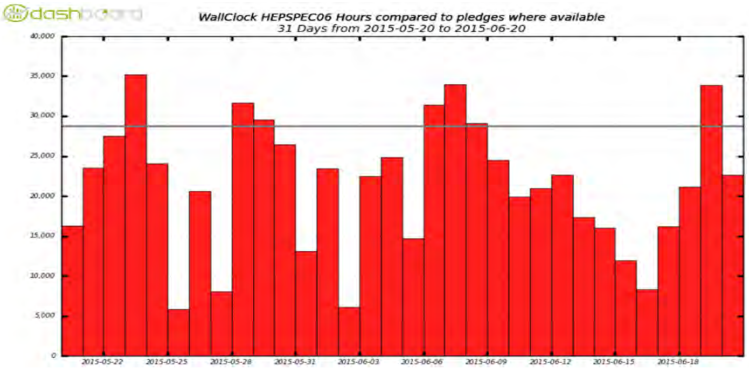
X = 19700 S = 6700

WallClock HEPSPEC06 (simulated)



X = 22000 S = 6400

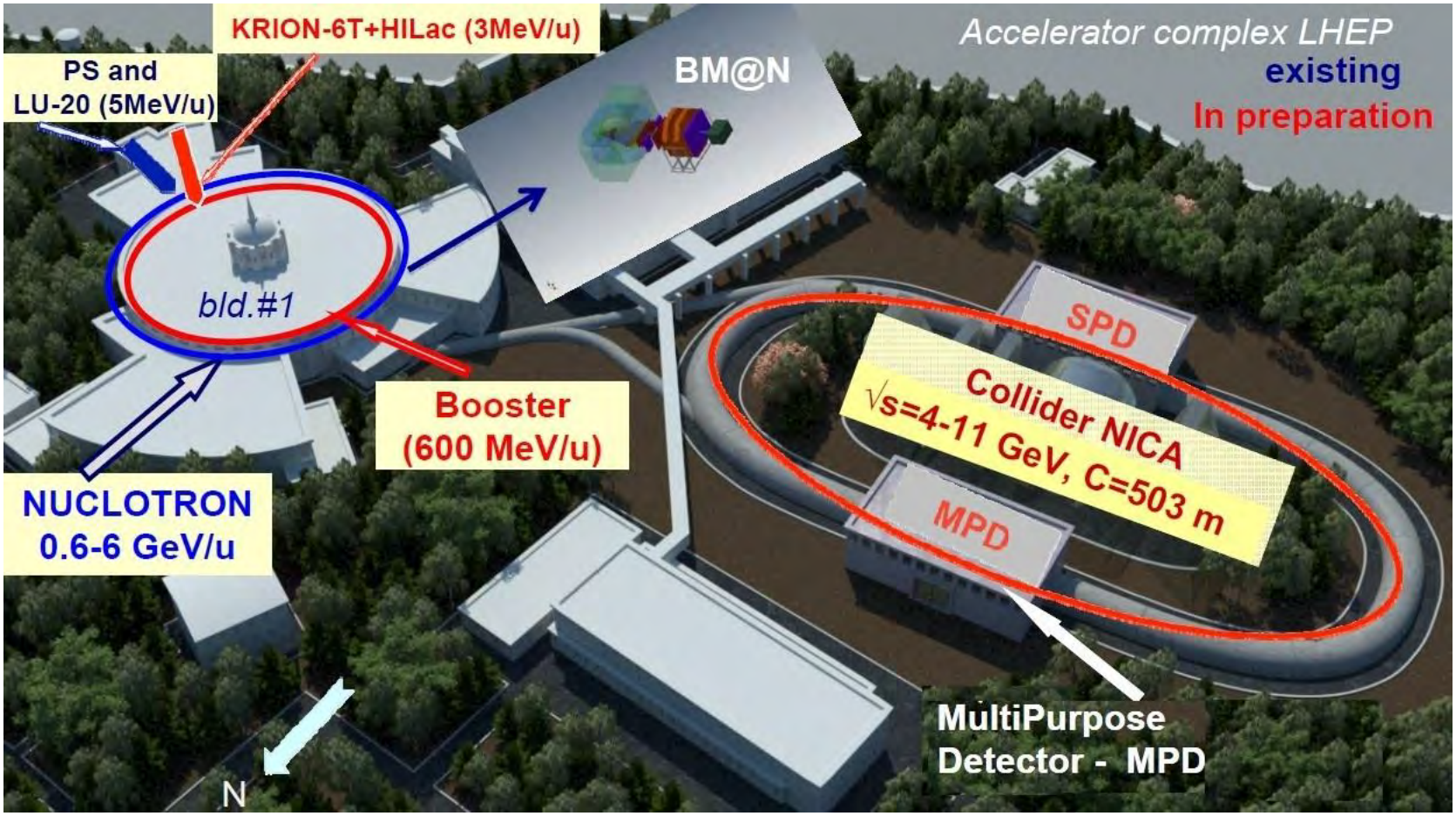
WallClock HEPSPEC06 (real)



X = 21300 S = 8100

These two examples among some others were used for the positive validation of the running CMS T1 model and encouraged us to simulate the more sophisticate T0/T1 system of NICA project.

NICA-MPD-SPD- BM@N



General view of the NICA complex with the collider and experiments MPD, SPD, BM@N

Implementation period (2016-2020)

- **2017** – first run of **BM@N experiment (1 PB / run)**;
- **2019** – initial functioning configuration of the NICA complex;
- **2020** – commissioning of the MPD detector and construction of the **SPD detector**;
- **2021** – first run of the experiment using the **MPD detector (10 PB / run)**

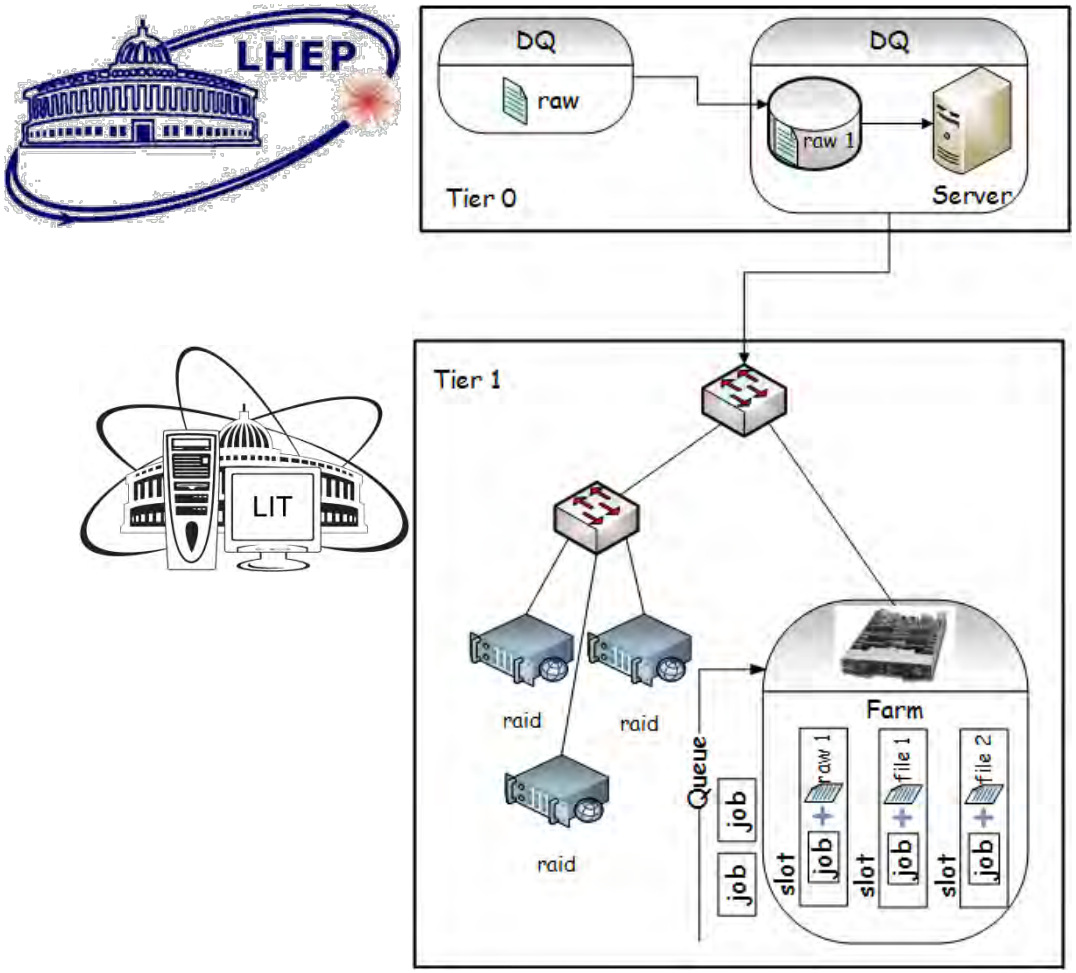
General goal:

to develop a distributed data processing system for the NICA complex

Goal for 2017:

to develop a data processing system for BM@N experiment

General scheme of the BM@N Data Processing System



BM@N Data processing characteristics

Characteristic	Value
The frequency of events occurrence at the output of the electronics	**10000 Hz**
The size of an event at the output of the electronics	**1 MB**
The number of events in a file	**5000 events**
The number of events in one **express analysis** job	**250 events**
The processing time of one event	**1 s**

SyMSim is used to choose a proper architecture of the BM@N computing system infrastructure.

Expected run of BM@N experiment

Running time – *1500 hours*.

*1000 cores are used **for processing***.

1. The first 120 hours – **adjusting**, without recording information.

2. 1000 hours – **data accumulation**, recording information (**flow 70%** of max).

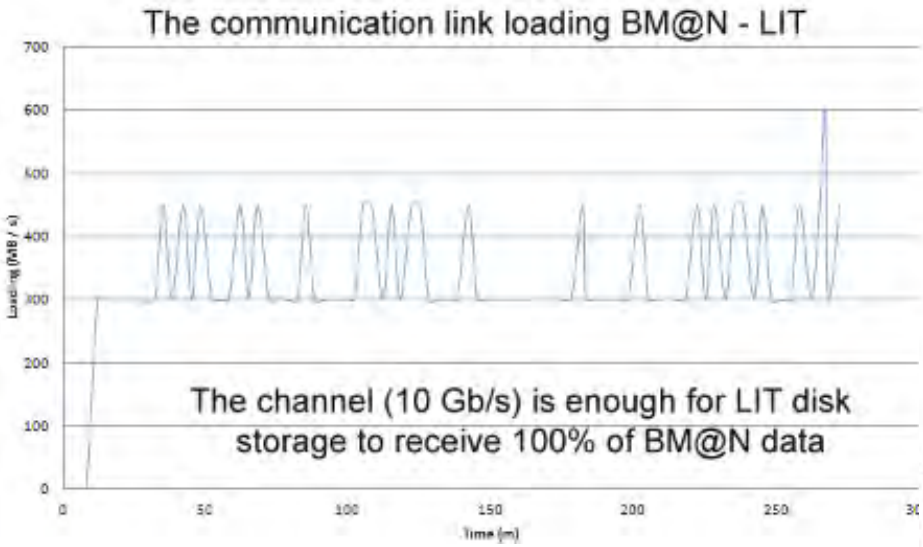
Express processing - ? cores.

Full processing - ? cores.

3. 380 hours – **data accumulation (flow 100%)**.

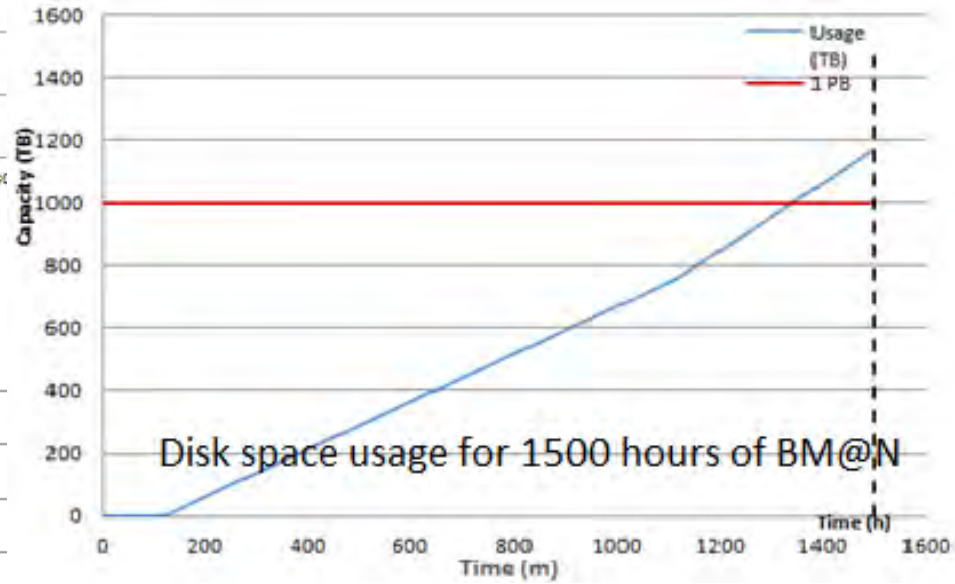
4. After data accumulation completion – *full processing 1000 cores*.

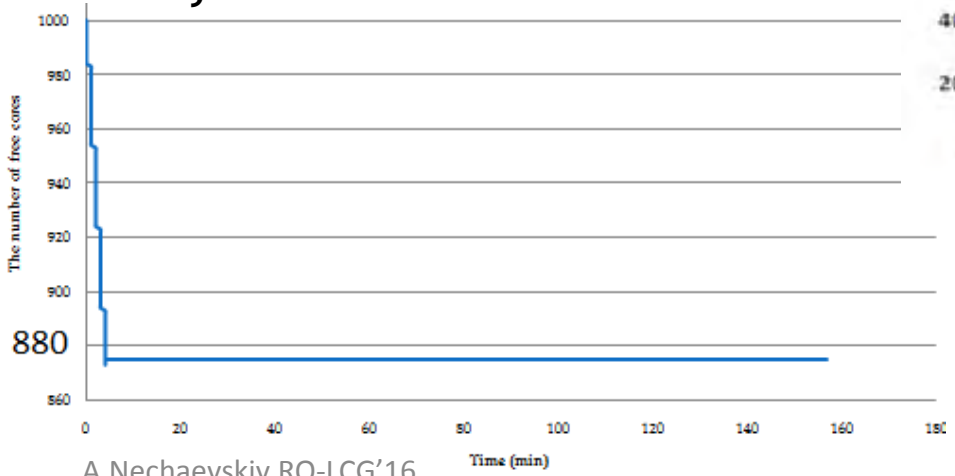
BM@N Data processing simulation



600 MB/s is a maximum load applied to the communication channel between BM@N and LIT.

1200 TB of disk capacity available for storing all files;

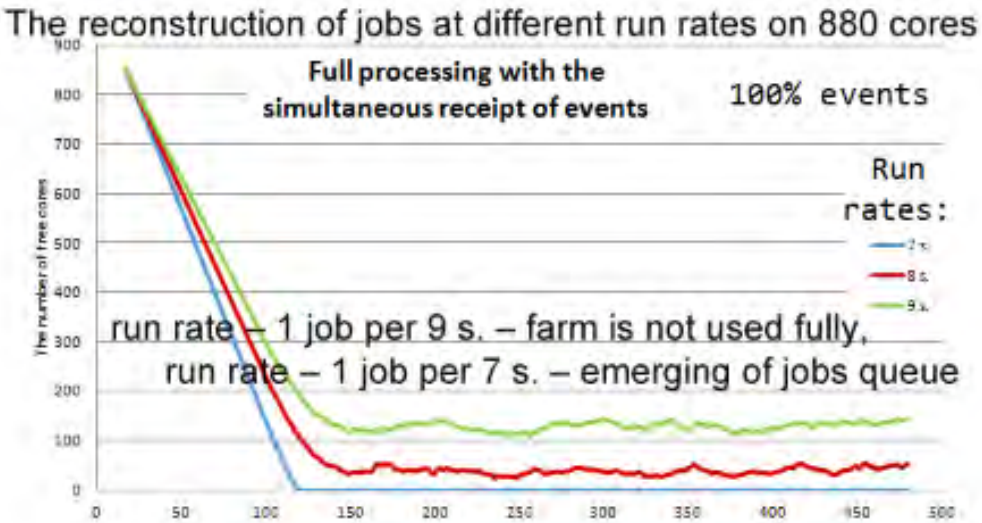




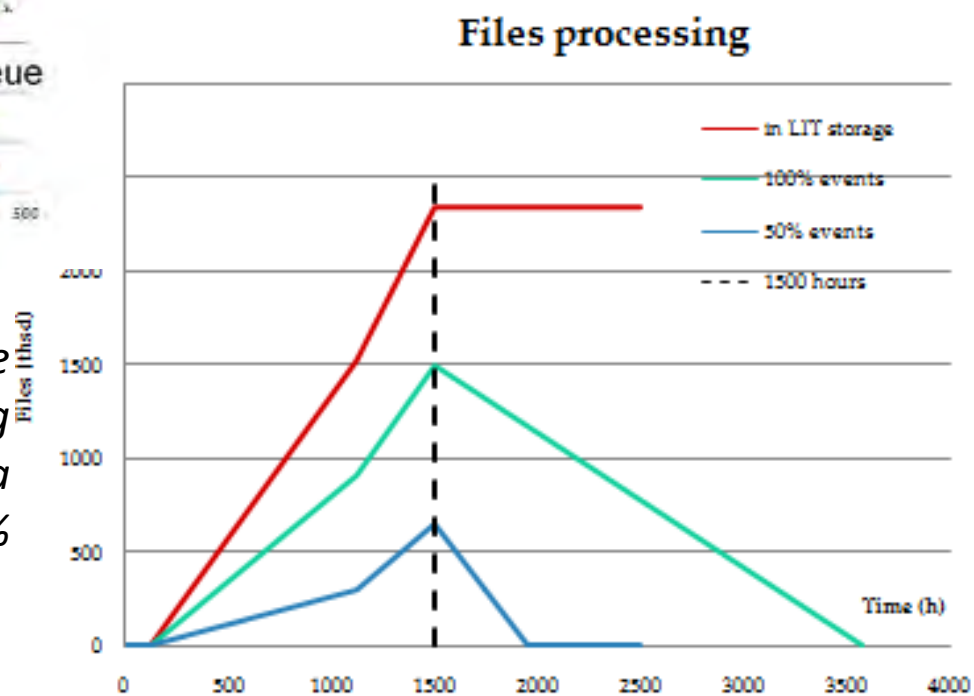
880 cores sufficient for full processing (if the rate of data accumulation is constant)

BM@N Data processing simulation

On-line event processing and data accumulation are done simultaneously

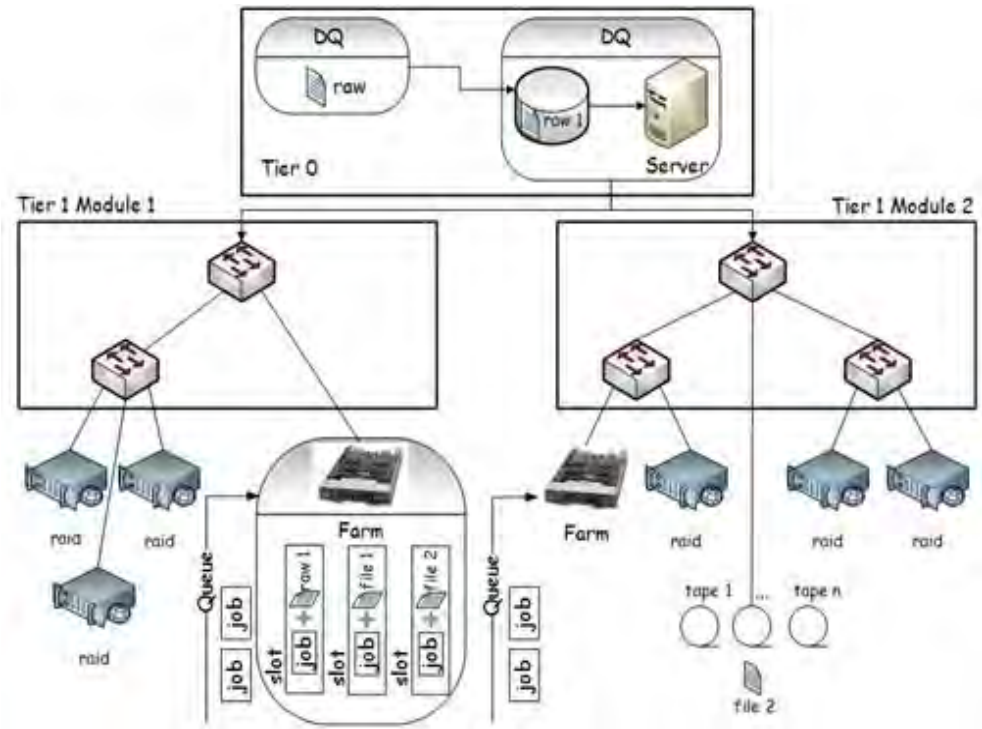


Run rate job / 8 s. allows to avoid low usage of the farm resources and emerging jobs queue.



After the experiment run is completed, the scenario with on-line 100% events processing on 1000 cores requires 2000 hours of extra time, while the second scenario with 50% events would require 500 hours only.

Simulation evolution: from CMS Tier1 to NICA Tier0-Tier1



Tier 0 module denotes the center of data gathering from the experiment (either MPD or SPD). Obtained raw data are to be stored on disks. One of planned problems is to recommend the volume of the disk store and a temp of data transfer to the robotized library which is the part of Tier 1 center. This two-level structure is interconnected by a local area network

DQ on this scheme denotes not only DAQ of the corresponding experiment, but includes also the means of communications and buffer cleaning. (LAN).

Data storage and processing scheme of Tier0-Tier1 level

Initial information to start simulation are **parameters of**

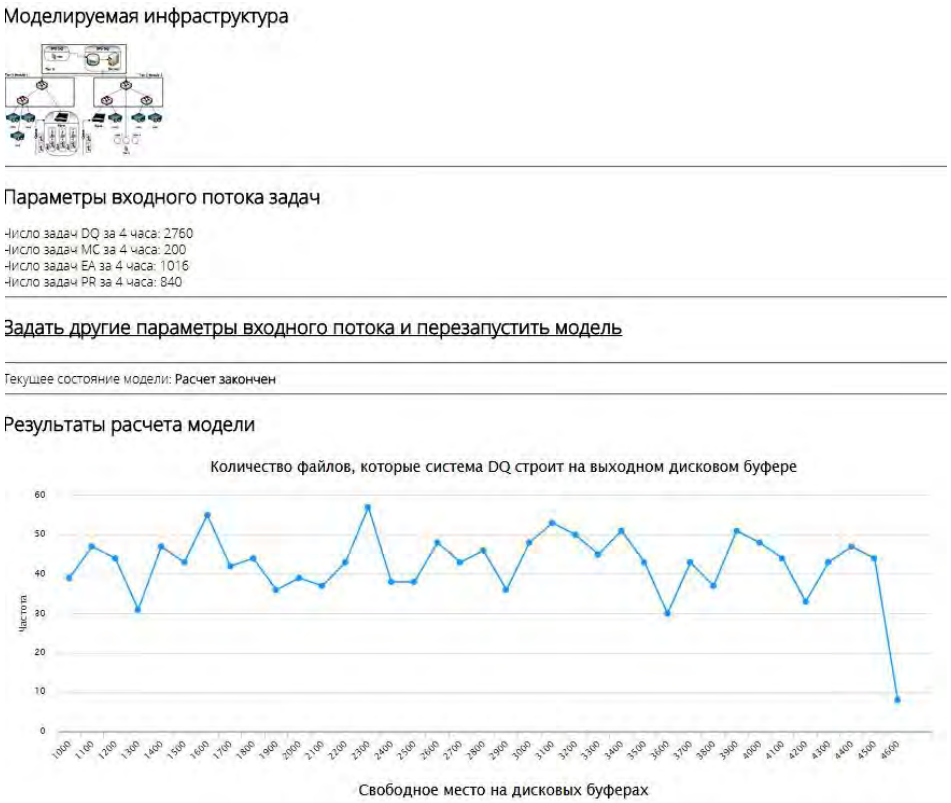
- **setup of designed hardware**
- **data flow,**
- **job stream**

} **their characteristics are taken from Real data of CMS Tier1 monitoring and TDR DAQ MPD**

Simulation of T0/T1

Web-portal functions

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

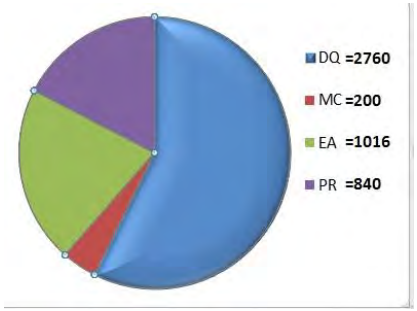


Simulation algorithm is designed that at the initial time all buffers are empty, the processor is not loaded and data are not transferred. **Therefore the initial transition process must be excluded from the analysis.** It also happens when the current job flow stops.
The result of the simulation program is a sequence of records in the database, which reflects all the events occurring at the system.

Examples of simulation results 1

Example 1

Estimated rate of NICA-MPD project experimental data to be transferred to Tier 1 data center is about 24 PB by one month of the MPD detector work



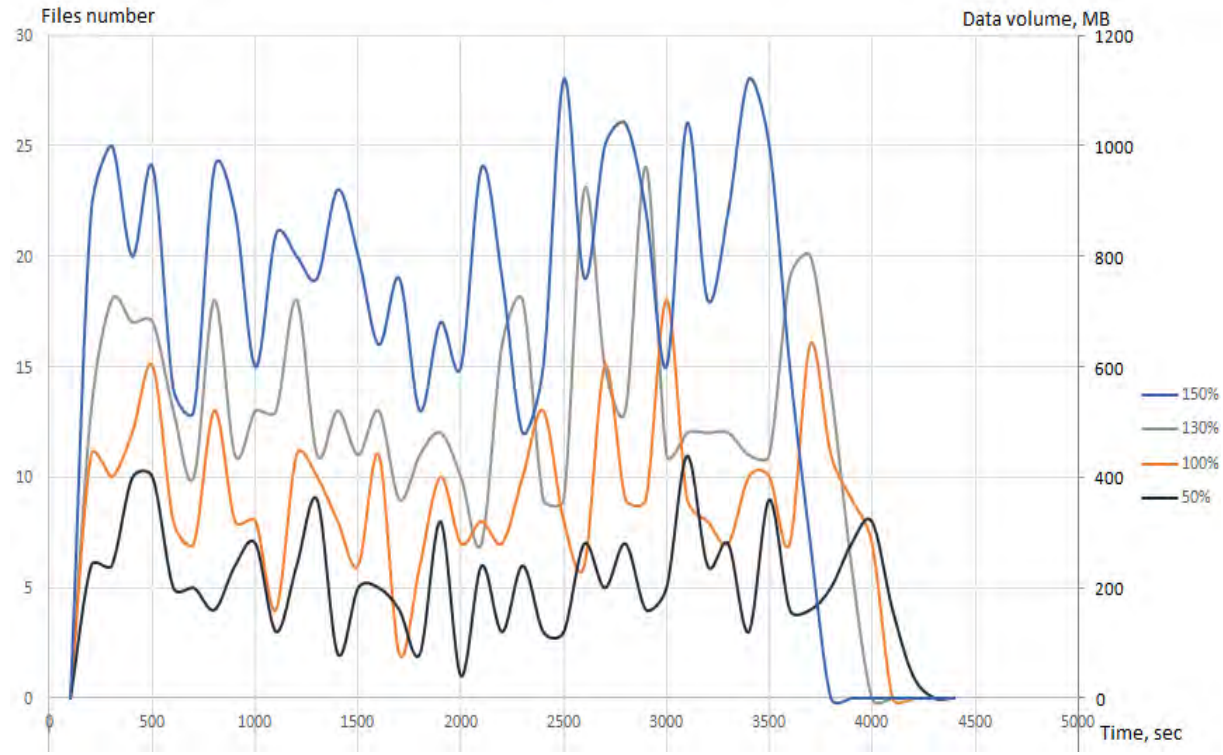


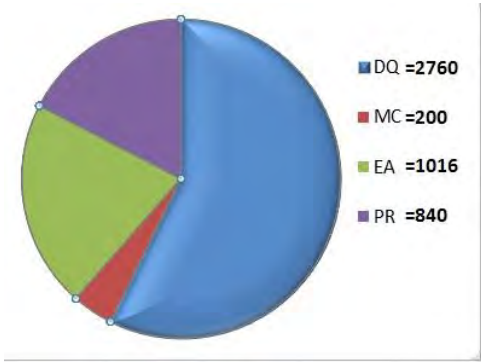
Fig.1 Number of DAQ data files stored on output disk buffer for growing data volumes

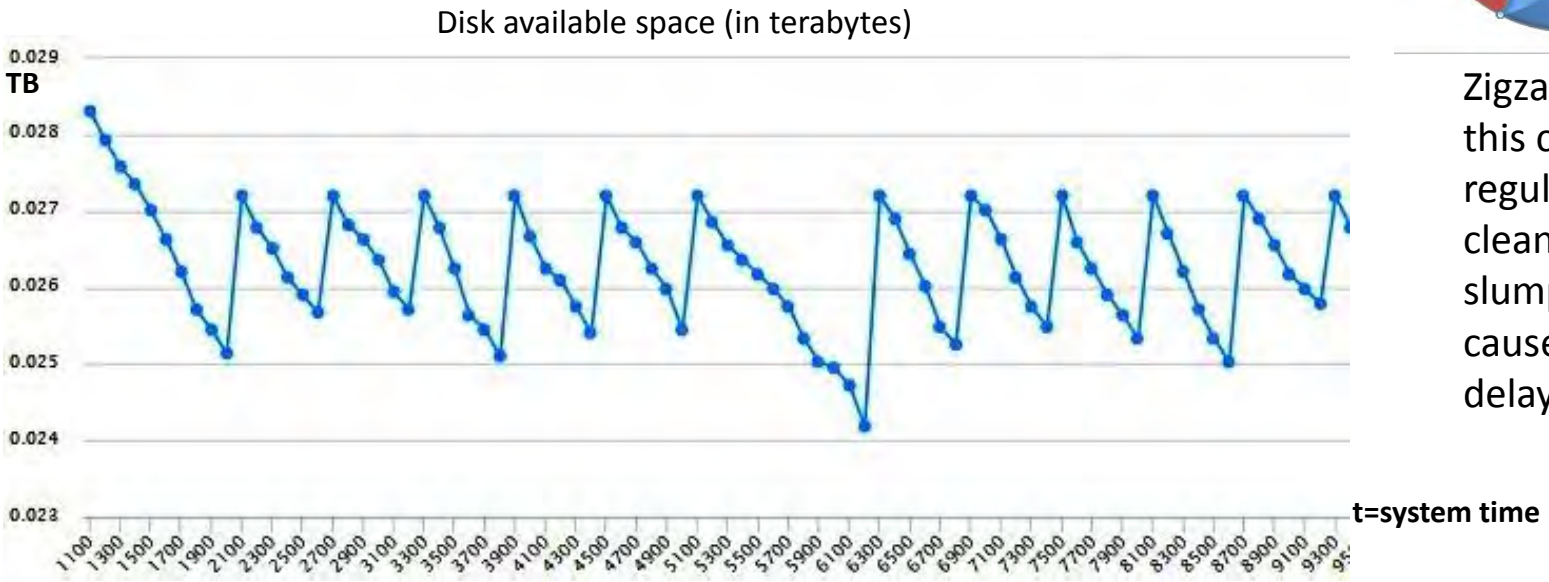
Simulation result shows what happened in the grid/cloud system if the data volumes are grow up to 1,5 times for example. This simulation result allows one to understand how the intensity of the input stream determines the reserves of the system capacity

Examples of simulation results 2

Example 2

What buffer size is needed to store input files on tapes without losses





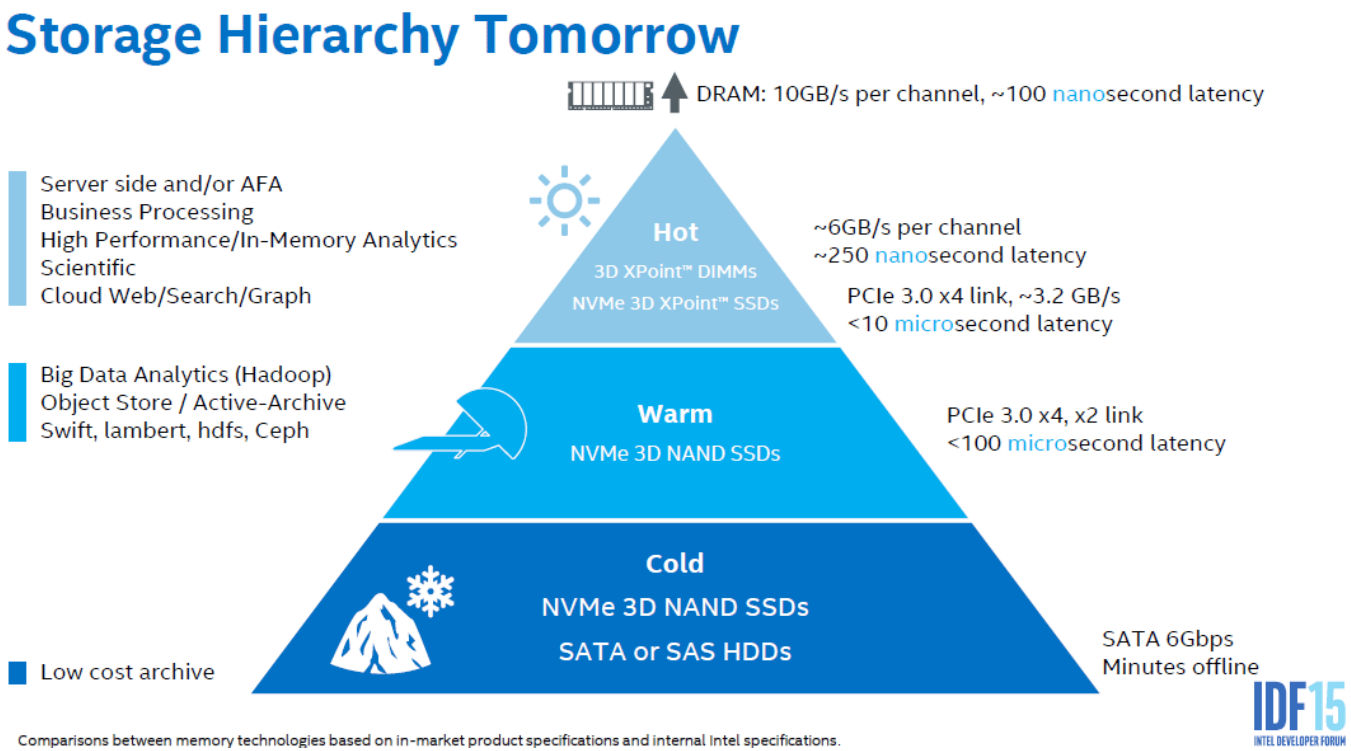
Zigzag shape of this curve is due to regular buffer cleaning. The sharp slump in the middle is caused by end-of-tape delay

Results shows that due to clever buffer cleaning the buffer should not be too big, so we can place it in RAM operational memory.

Conclusions

1. The program SyMSim for simulation of grid-cloud structures is developed and tested on a simplified model of JINR Tier1 site.
2. The next simulation was accomplished for BM@N and MPD computing facilities of NICA complex. It confirms good potential of our simulation approach.
3. SyMSim structure is sufficiently general and flexible to allow to replace our present simplifications into more real conditions in future developments.
4. It can also be used to solve design problems and the subsequent development of data repositories, not limited to the physical experiments area.

Conclusions



1. The next generation of the storage systems, network technologies and others is coming soon
2. Creation of the MegaExperiments are going a several years.
3. The simulation system help us to understand what we need and what we can get in the years

Our Team

Ososkov G. A.
Dr. of Science in physics and mathematics, Professor;

Trofimov V.V.,
Leading programmer

Nechaevskiy A.V.,
Software engineer

Uzhinskiy A.V.,
Candidate of Science (PhD) in technology;
Leading programmer

Prjahina D.I.,
Software engineer

Thank you for the attention!

More about SyMSim:

1. *Simulation concept of NICA-MPD-SPD Tier0-Tier1 computing facilities* (2016) Physics of Particles and Nuclei Letters, 13 (5), pp. 693-699.
2. *The JINR Tier1 Site Simulation for Research and Development Purposes* (2016) EPJ Web of Conferences, 108, art. no. 02033
3. *Web-Service Development of the Grid-Cloud Simulation Tools* (2015) Procedia Computer Science, 66, pp. 533-539.
4. *Synthesis of the simulation and monitoring processes for the development of big data storage and processing facilities in physical experiments* (2015) Computer Research and Modeling, T7, №3
5. *Grid and cloud services simulation as an important step of their development* (2015) Systems and Means of Informatics, Volume 25, Issue 1

symsim.jinr.ru symsim@jinr.ru