Digital twins of distributed data acquisition, storage and processing centers: status and prospects

D. PRIAKHINA, V. KORENKOV, V. TROFIMOV

Meshcheryakov Laboratory of Information Technologies, JINR

10.07.2025

Distributed data acquisition, storage and processing centers (DDC)

WLCG Worldwide LHC Computing Grid

MICC JINR MULTIFUNCTIONAL INFORMATION AND COMPUTING COMPLEX

https://micc.jinr.ru/

Data Center Institute of High Energy Physics Chinese Academy of Sciences

- Geographically distributed infrastructure.
- Designed to work with extremely large amounts of data.
- Consists of various types of resources.
- Collective shared access to data storage and processing resources.

Digital twin of DDC

Equipment (physical assets)

Network infrastructure

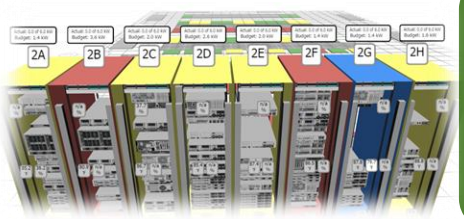
Engineering infrastructure

Data flows & job flows

- Improving the technical characteristics of computing systems.
- Making decisions on the choice of equipment configurations.
- Scaling and resource management.

etc…

Digital twin of DDC



Equipment (physical assets)

Network infrastructure

Sunbird

FORWARD NETWORKS

future facilities

Engineering infrastructure

Data flows & job flows

DDC simulation tools:

- Bricks, OptorSim, GridSim libraries;
- SyMSim software complex…

Digital twin (DT)

Real-time operation throughout the entire DDC life cycle.

COMPUTER MODEL

INPUT DATA

- Architecture and hardware parameters of DDC.
- **Characteristics of data flows and job flows**.
- **Parameters of random processes of the system functioning** (data loss, failures, changes in the equipment performance, etc.).

FUNCTIONAL PURPOSE

- Designing of DDC.
- Analysis of the efficiency and reliability of DDC.
- Testing scaling scenarios based on **data flows and job flows requirements**.
- Assessment of the required amount of resources for specific tasks.
- Checking job flows management strategies.

Method of creating digital twins of DDCs

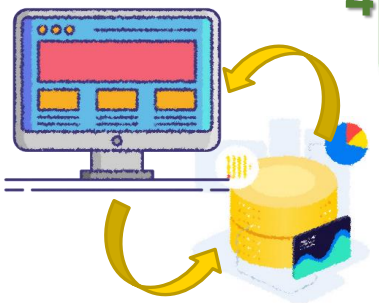
1. Obtaining data about the DDC: equipment parameters, data flows and job flows; purpose of building a DT; probabilistic characteristics of processes in the DDC
2. Description of the DDC structure and connections between the components
3. Formation of a DDC virtual image based on the data and the description of its structure
4. Simulation of the DDC
5. Graphical representation of the DT results

Method of creating digital twins of DDCs

1. Obtaining data about the DDC: equipment parameters, data flows and job flows; purpose of building a DT; probabilistic characteristics of processes in the DDC
2. Description of the DDC structure and connections between the components
3. Formation of a DDC virtual image based on the data and the description of its structure
4. Simulation of the DDC
5. Graphical representation of the DT results

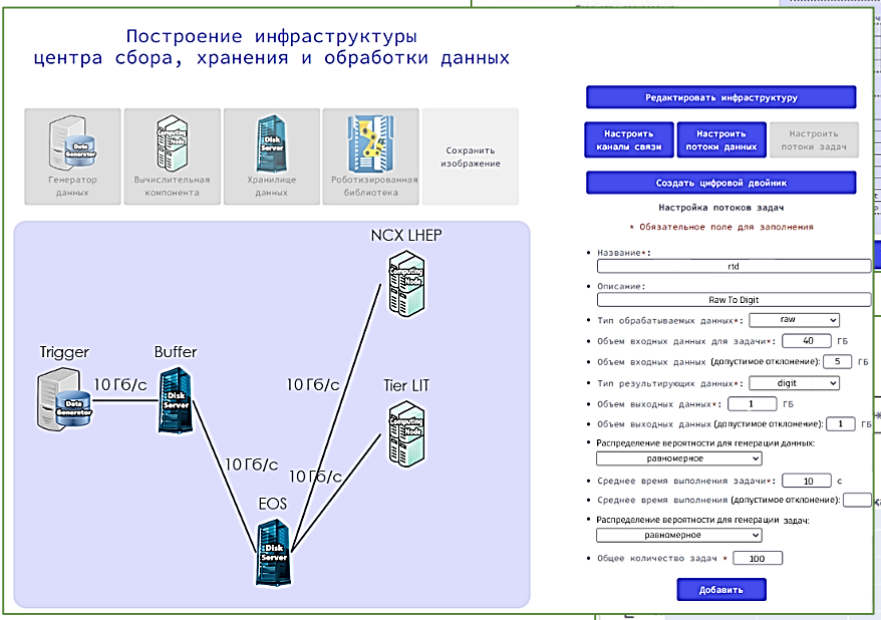
DT's Core

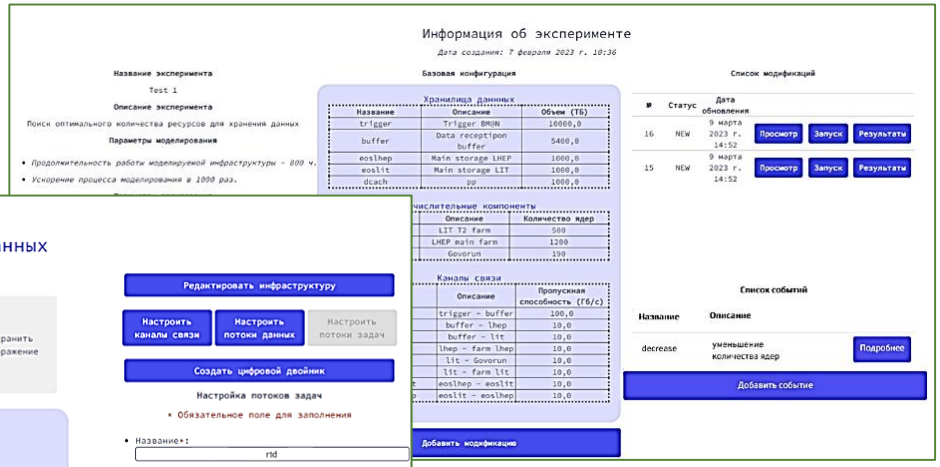
User Interface



User Interface functionality

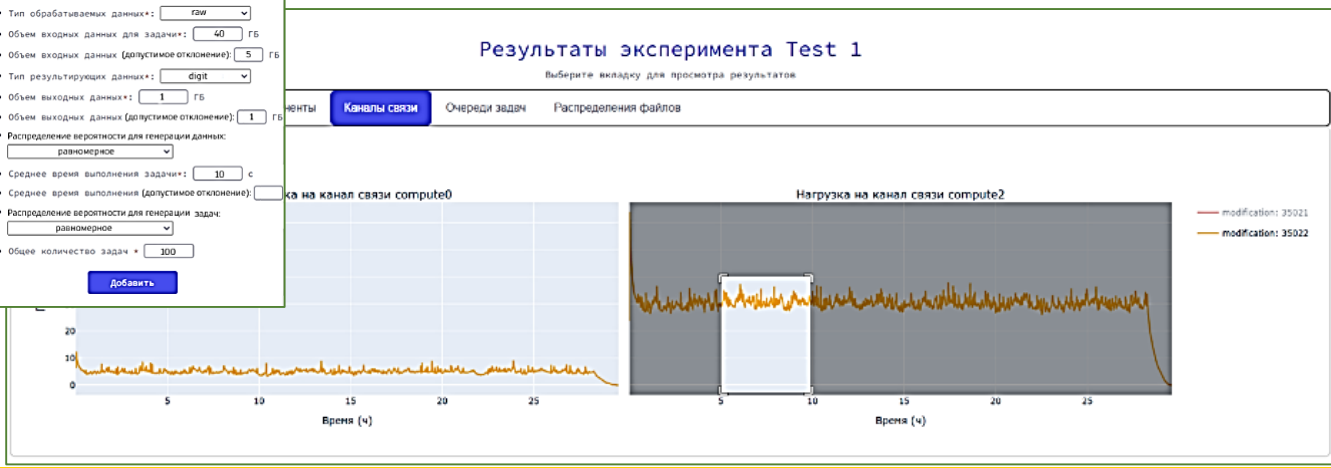
- Building the DDC structure.





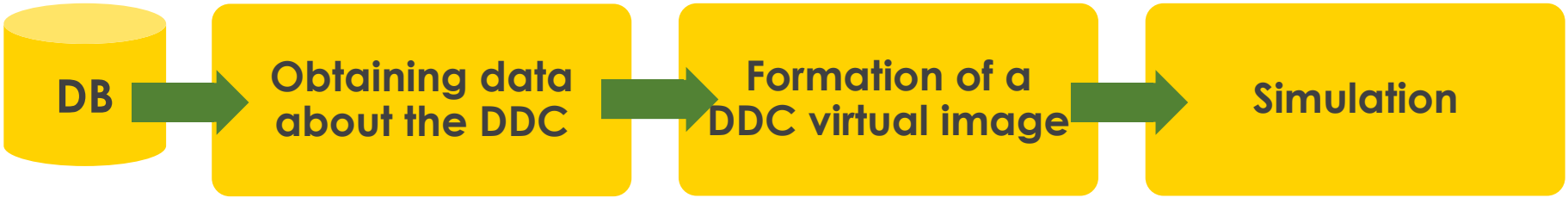
- Setting the parameters of the DDC equipment, data flows & job flows characteristics.
- Configuring DDC scaling scenarios.
- Launching the DT.

- Viewing the DT results.



DT's core

Ensures the functioning of the DT.



- **Probability distributions** are taken into account when forming data flows, job flows, and equipment performance criteria.

Generating parameters for data flows and job flows:

$$pr(x) = \frac{1}{b-a}$$

- the probability density function of a uniform distribution, where
[a, b) — the interval for changing the parameter value

$$pn(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- the probability density function of a normal distribution, where
μ — the average value of a random variable,
σ — standard deviation

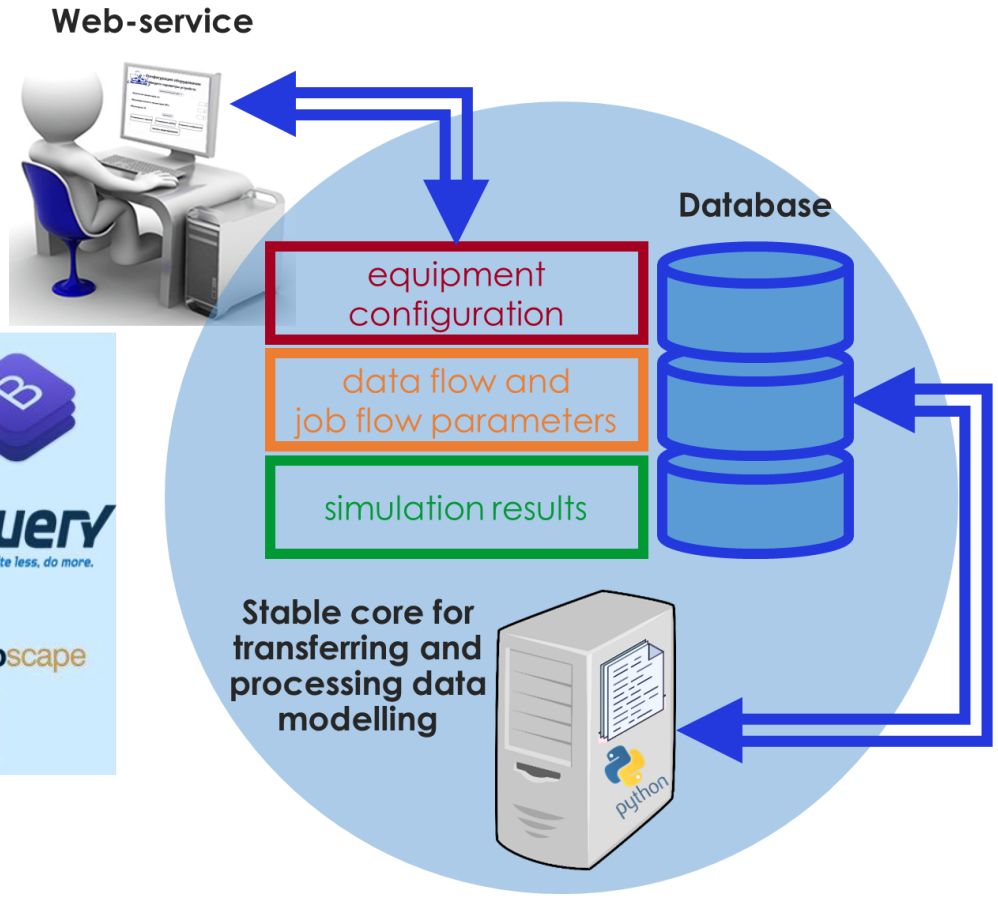
Event flow generation (exponential distribution):

$$\tau = -\frac{1}{\lambda} * \ln(r)$$

τ — the interval between random events,
λ — average number of events per unit of time,
r — an uniformly distributed random number [0;1]

Software complex for creating digital twins of DDC

Technology stack



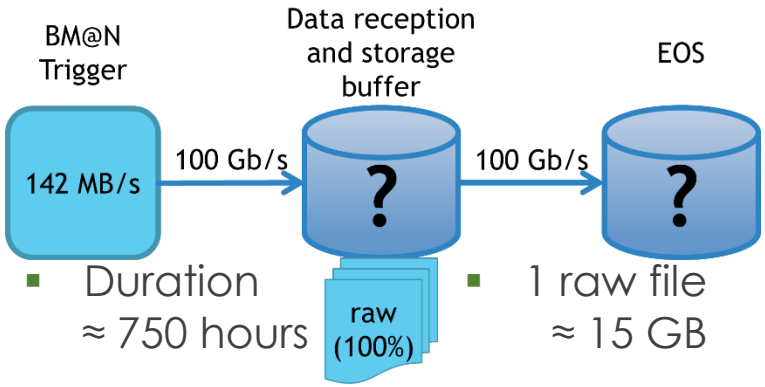
Digital twins of DDC for mega science experiments

Russian Federation, Moscow region, Dubna, Joint Institute for Nuclear Research

Verification of the DT's core

Initial data: The results of monitoring the computing infrastructure of the BM@N experiment of the NICA complex (physical session 2022-2023)

Acquisition and storage of experimental data



- Duration ≈ 750 hours
- 1 raw file ≈ 15 GB

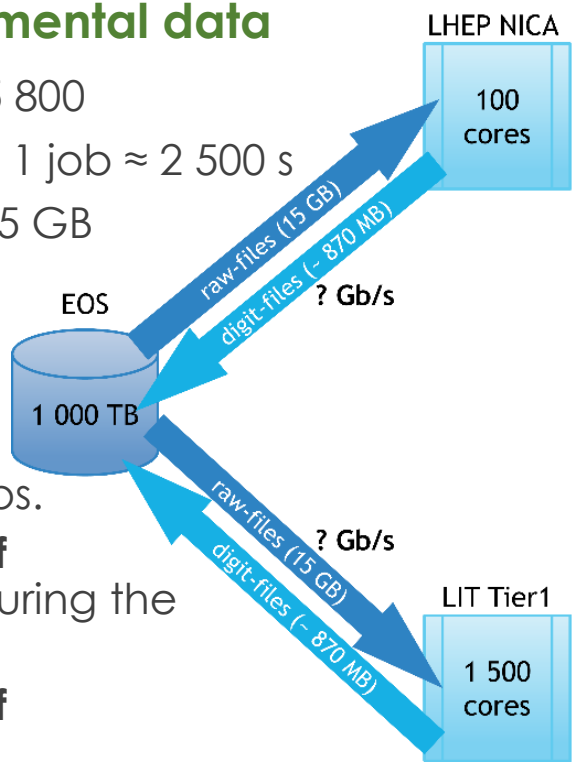
Tasks

1. Find the **amount of resources** that are needed **to store all raw-data**.

2. Find the **number of raw files** in the EOS storage.

Executing jobs of conversion of experimental data

- Number of jobs – 25 800
- Time to complete a 1 job ≈ 2 500 s
- 1 raw (input) file ≈ 15 GB
- 1 digit (output) file ≈ 870 MB



Tasks

1. Find the **total execution time** of all jobs.

2. Calculate the **load of computing resources** during the execution of jobs.

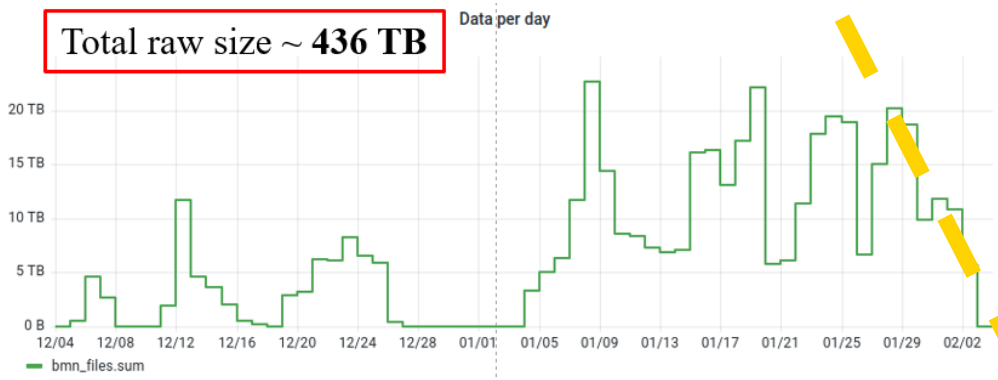
3. Calculate the **load of communication links**.

Verification of the DT's core

Acquisition and storage of experimental data

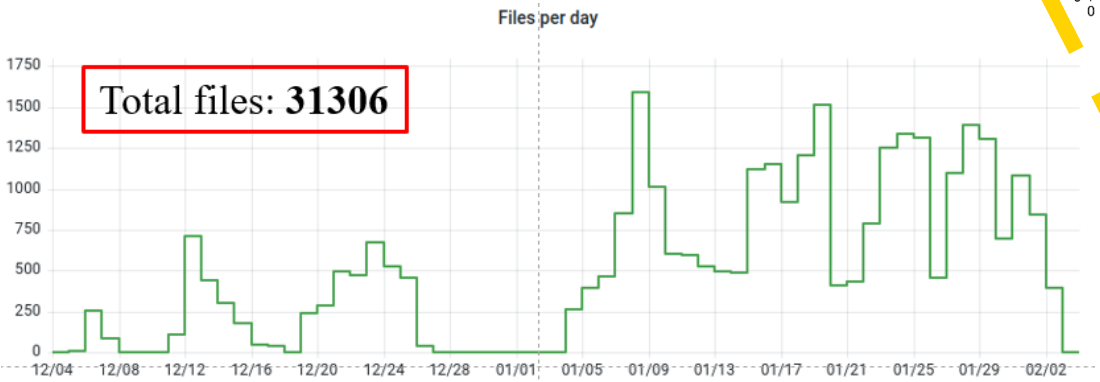
Monitoring

Total raw size ~ **436 TB**



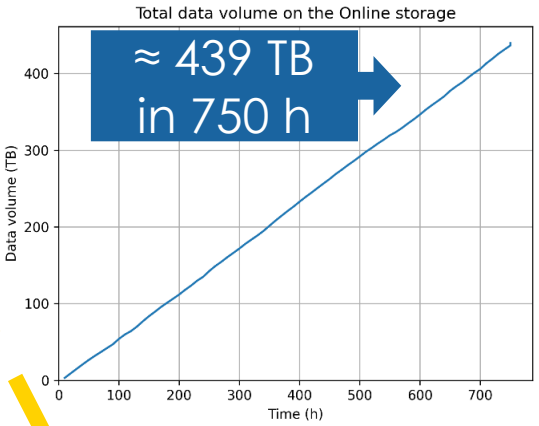
- Nonuniform distribution.

Total files: **31306**



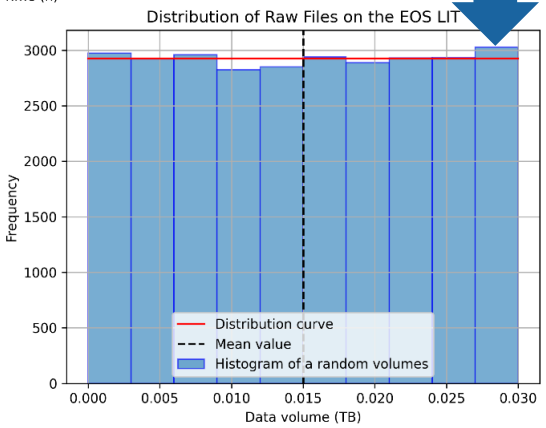
- Data is missing in some time periods.

Digital twin



Data generation efficiency: 20%

≈ 29 241 raw files

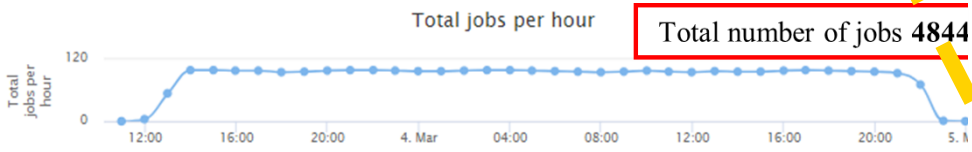


Verification of the DT's core

Computing resource usage and jobs execution time

Monitoring

LHEP NICA (100 cores)

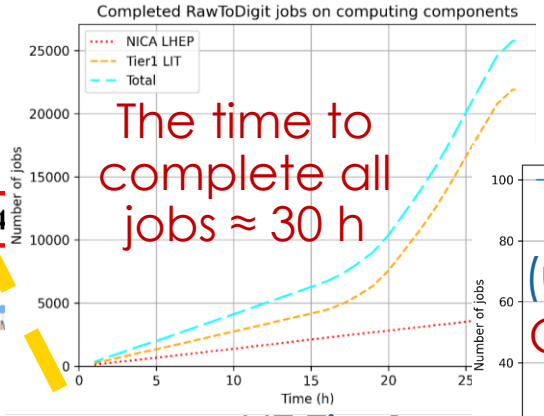


- Uniform use of resources ≈ 100 jobs / h
- 19% jobs of the total number

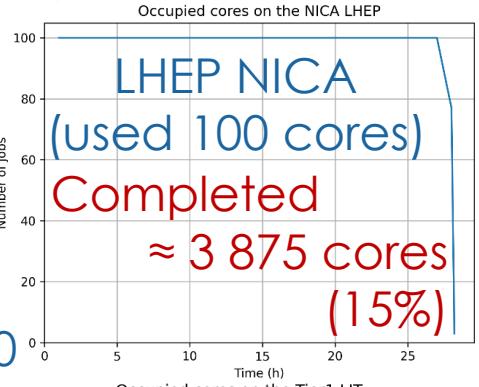
LIT Tier1 (1500 cores)

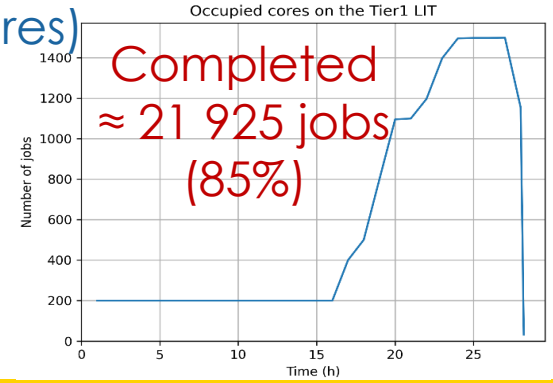
- Ununiform use of resources
- 200 – 1 500 jobs / h
- 81% jobs of the total number

Digital twin



LIT Tier1 (used from 200 to 1500 cores)

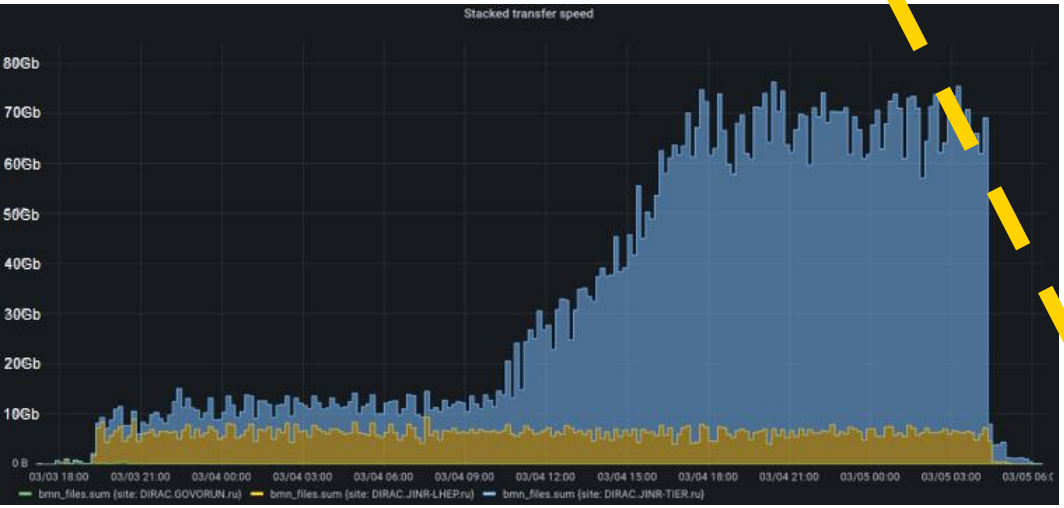




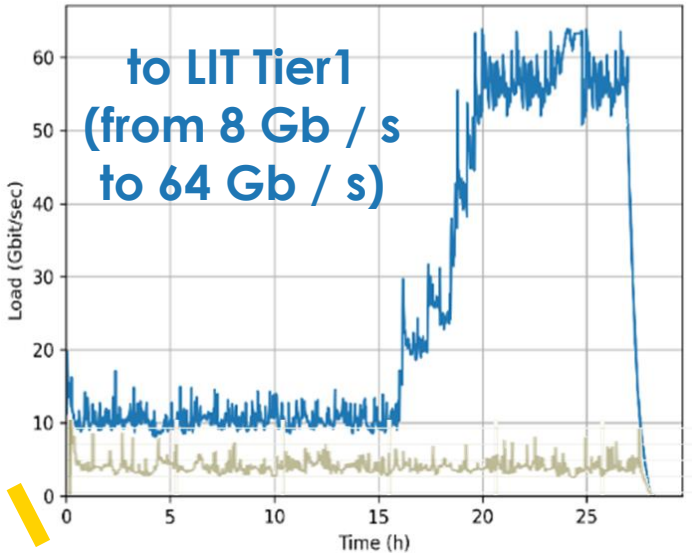
Verification of the DT's core

Computing resource usage and jobs execution time

Monitoring | **Digital twin**



- Data transfer rate between to LHEP NICA ≈ 4 Gb / s
- Data transfer rate between to LIT Tier1 from 8 Gb / s to 64 Gb / s



до LHEP NICA (≈ 5 Gb / s)

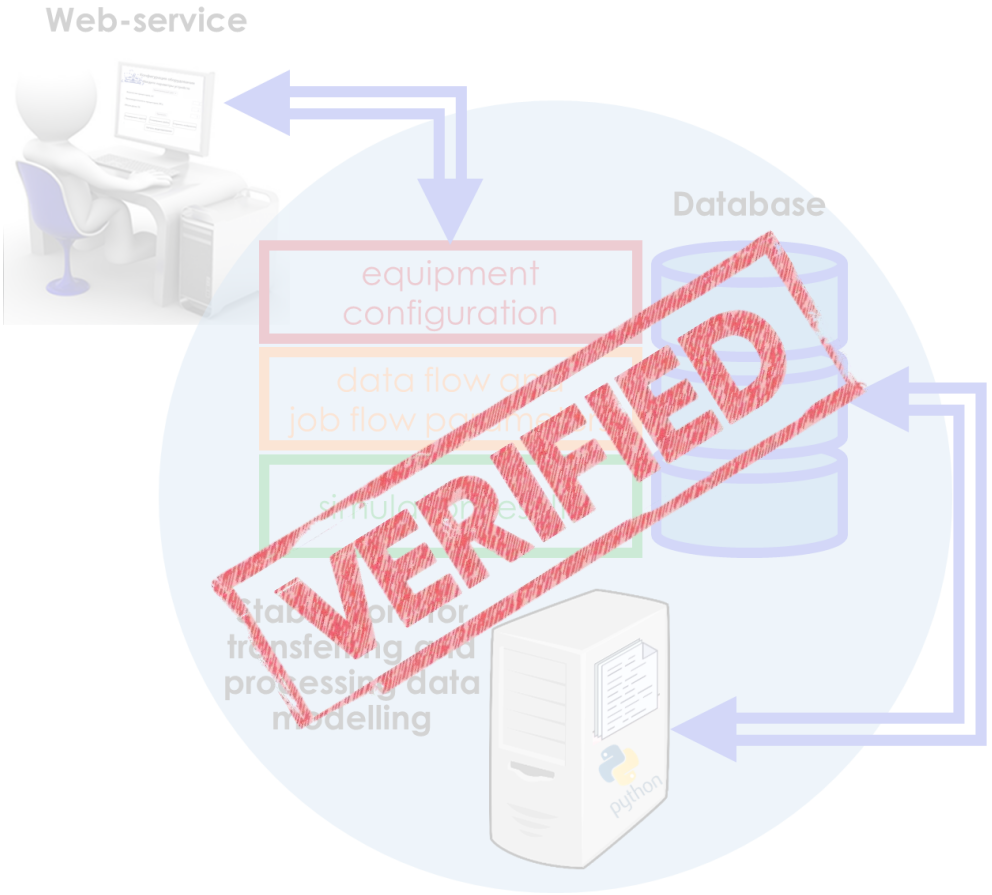
Verification of the DT's core

Conclusions

1. **A series of experiments** was conducted.

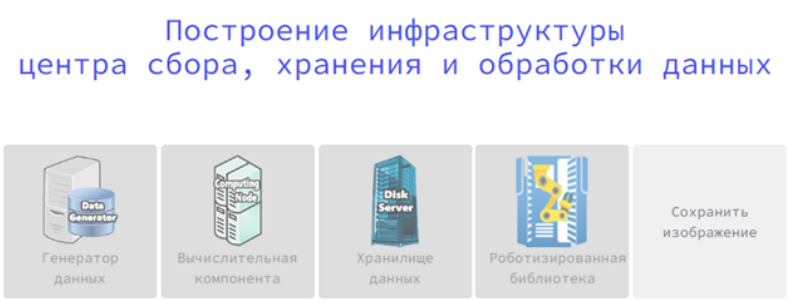
2. DT results **do not exceed three standard deviations** from the average value obtained by monitoring results.

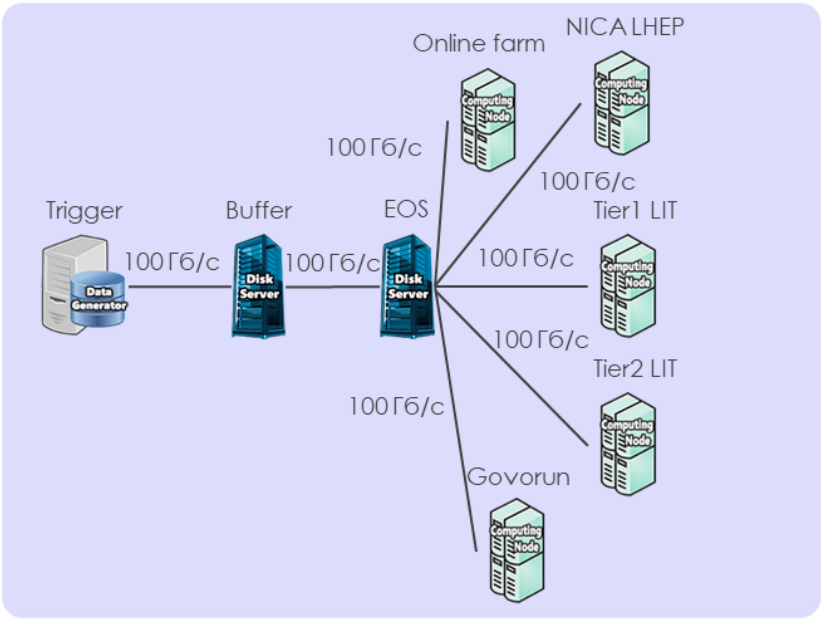
3. **The value of accuracy is sufficient for further use** and creation digital twins of DDCs, which will be used to solve the problems of design and development the computing infrastructure of scientific megasciences experiments.



Application of the software complex to create the DT

The DT of the computing infrastructure of the BM@N experiment





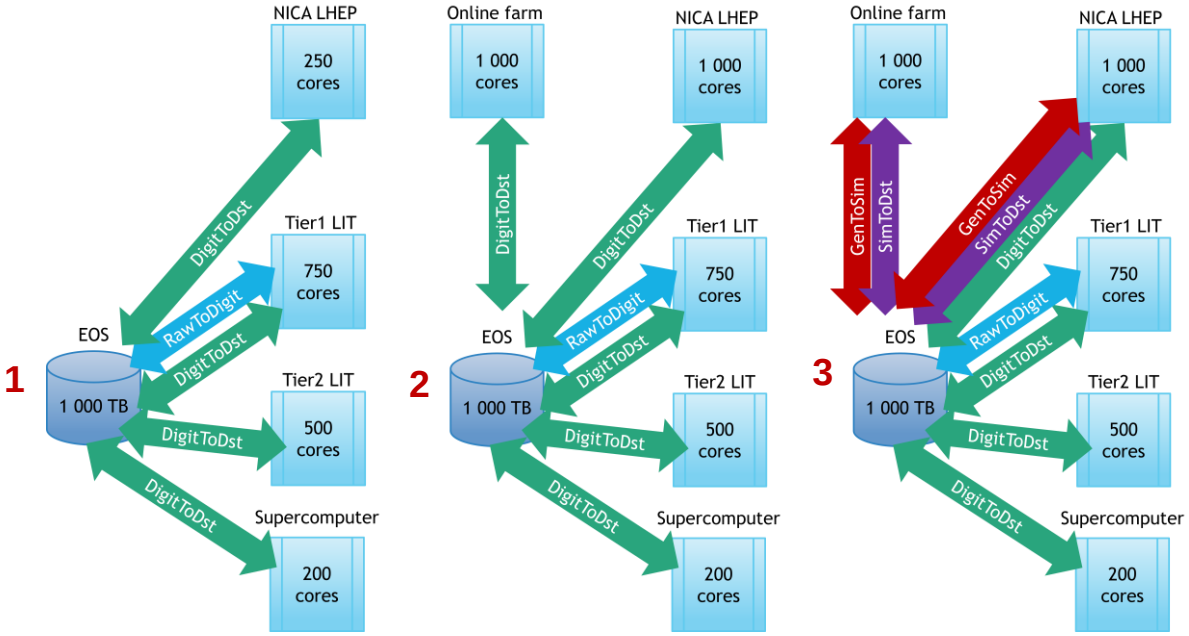
Goals

- **to estimate the resource requirements for data storage and processing** taking into account the planned parameters of data flows of future experiment sessions;
- to get an approximate **time it will take to process the data** at the end of the experiment session.

Application of the software complex to create the DT

The DT of the computing infrastructure of the BM@N experiment

Comparison of infrastructure configurations for data processing



1→2:
– Increased the number of resources on **NICA LHEP from 250 to 1 000 cores.**
– Added resources to **Online farm (1 000 cores).**
3: Added job flows for **simulation data processing**.

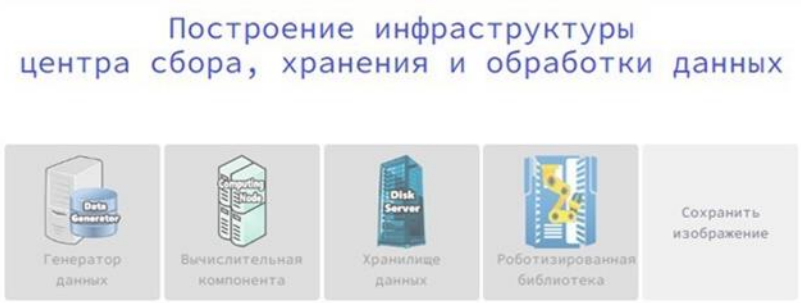
100% resource usage

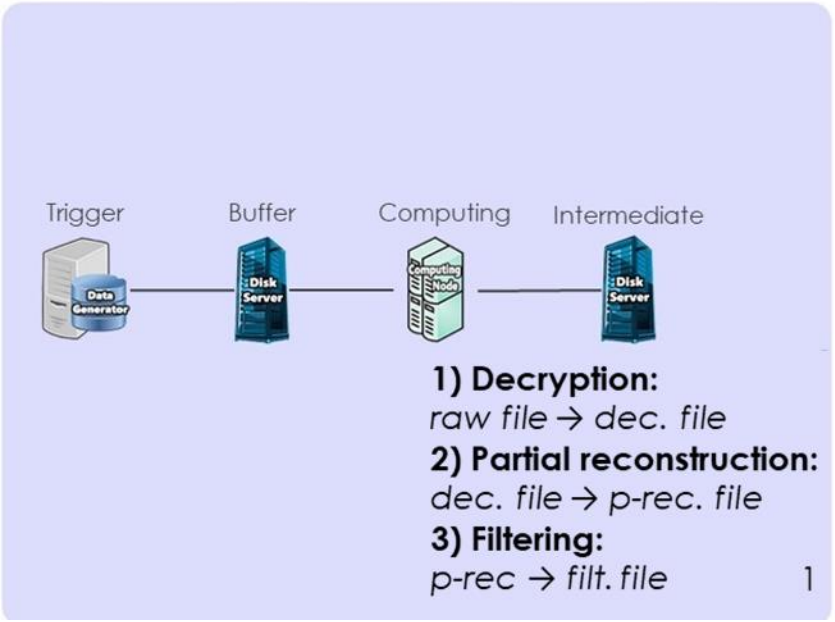
The process of converting raw-data **speed up by 2 times**

Results	1	2	3
Convert all raw-data to digit format(RawToDigit)	60 h (2,5 days)	**28 h** (1 day)	**28 h** (1 day)
The time of complete processing of all experimental data	432 h (18 days)	**240 h** (10 days)	**367 h** (15 days)
The time of complete processing of all simulation data	---	---	**130 h** (6 days)

Application of the software complex to create the DT

The DT of the computing system of the online data filter of the SPD experiment





Goals

- **design an efficient computing system** to ensure fast event reconstruction and real-time data filtering;
- **find out** the amount of data storage, network load, required amount of computing resources, etc.;
- **provide opportunities for further development** and optimization of the designed system.

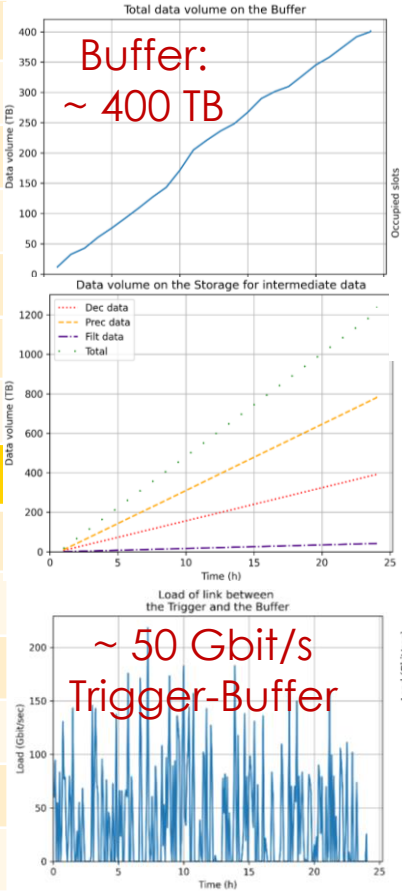
Application of the software complex to create the DT

The DT of the computing system of the online data filter of the SPD experiment

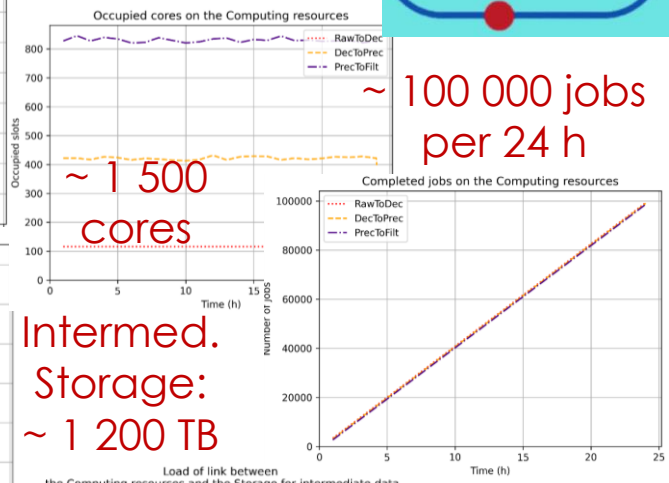
Initial data			
Facility operating time	24 h		
Data generation rate	20 GB/s		
Data generation efficiency	20%		
Event processing speed	1 000 events/s		
Number of data processing jobs	100 000		
raw-file	4 GB	**prec**-file	8 GB
dec-file	4 GB	**filtred**-file	450 MB

Results	
Buffer for storing raw-data	400 TB
Intermediate storage	1 200 TB
Computing resources	1 500 cores
Trigger-Buffer Comm. Link	50 Gb/s
Buffer-Comput. Comm. Link	40 Gb/s
Comput.-Storage Comm. Link	250 Gb/s

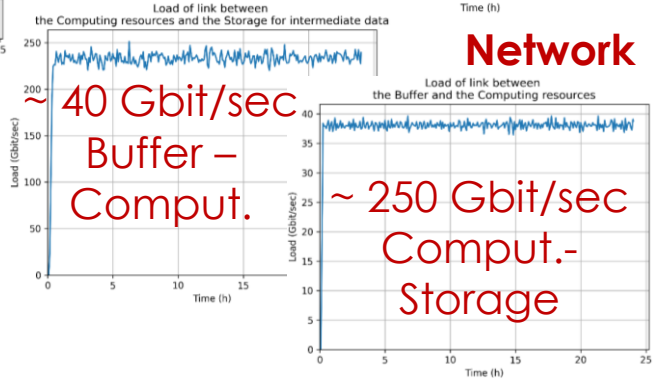
Data storages



Computing resources



Network



Current status

1. A new method of creating and using DTs is developed.

The method differs in the ability to simulate such processes as data storage and processing, taking into account the characteristics of data flows and jobs, the probabilities of failures and changes in the equipment performance and other processes.

2. The software complex has been developed to implement the method.

Software complex used to make decisions on the choice of the configuration of the DDC equipment according to the specified requirements.

3. The verification was carried out.

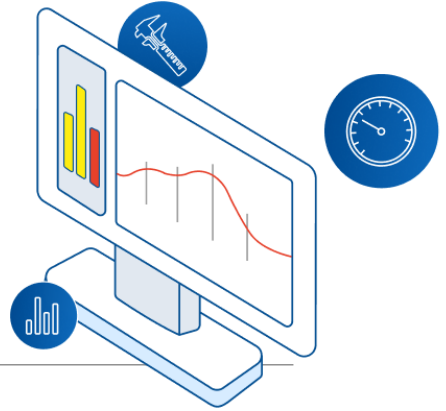
The adequacy of the constructed methods and algorithms was confirmed using the example of the computing infrastructure of the existing experiment.

4. The certificate of state registration of the computer program is received.

5. Experimental operation was carried out:

- the problem of finding an equipment configuration for the data acquisition, storage and processing system of the BM@N experiment of the NICA complex at the JINR is solved;
- the results of the research are used in the design of the computing system for the online data filter of the SPD experiment of the NICA complex at JINR.

Prospects



- **Improvement** the developed method.
- **Adding additional hardware description parameters** and other components of the data center.
- **Adding a multi-criteria optimization function** when choosing an equipment configuration.

As criteria, both more detailed **technical and cost parameters** of the equipment will have to be taken into account.

- **Upgrade the web service** for user interaction with the DT to improve the user-friendliness of the software complex.
- **Using the software complex** in the tasks of designing, creating, supporting and developing DDCs for large scientific projects.

Joint work on using software complex to simulate the computing infrastructure of the MPD experiment

Thank you for the attention!

D. PRIAKHINA, V. KORENKOV, V. TROFIMOV

Meshcheryakov Laboratory of Information Technologies, JINR

10.07.2025