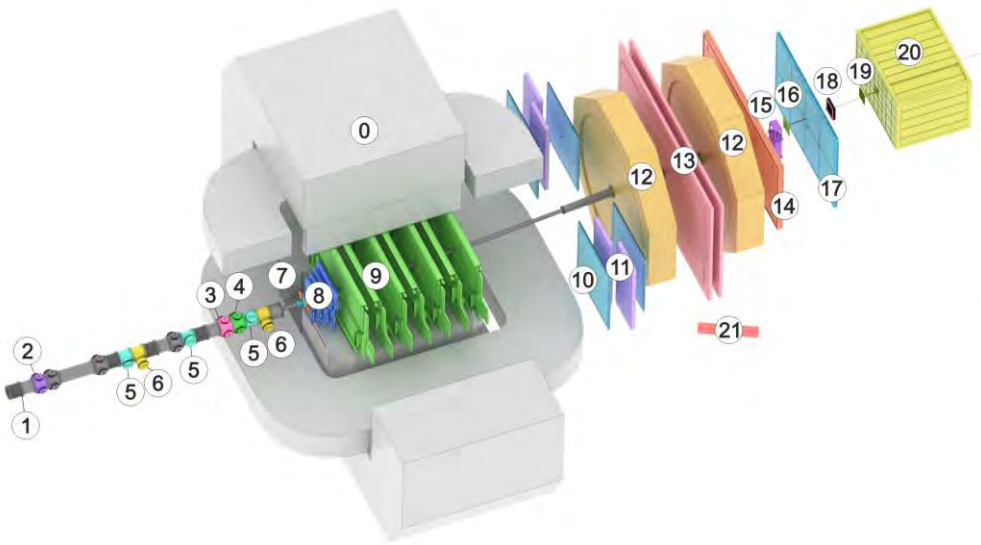
BM@N Run 8 raw data production on distributed infrastructure with DIRAC

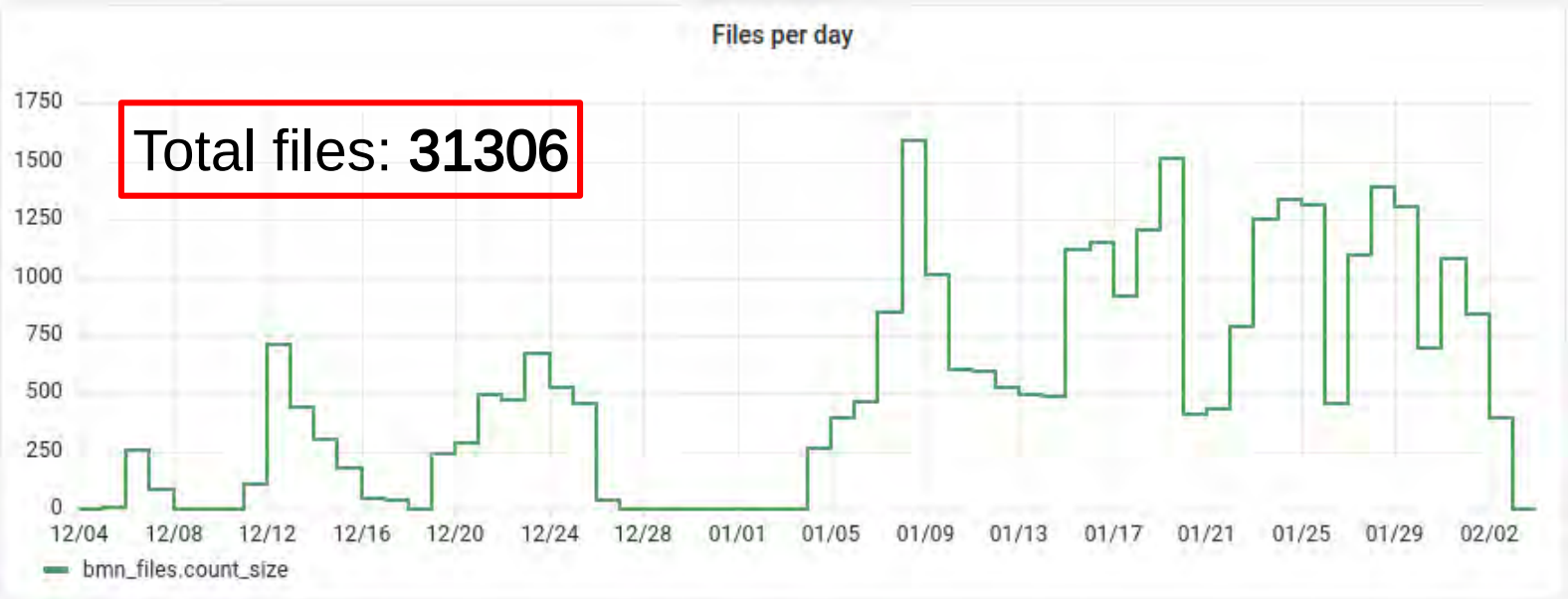


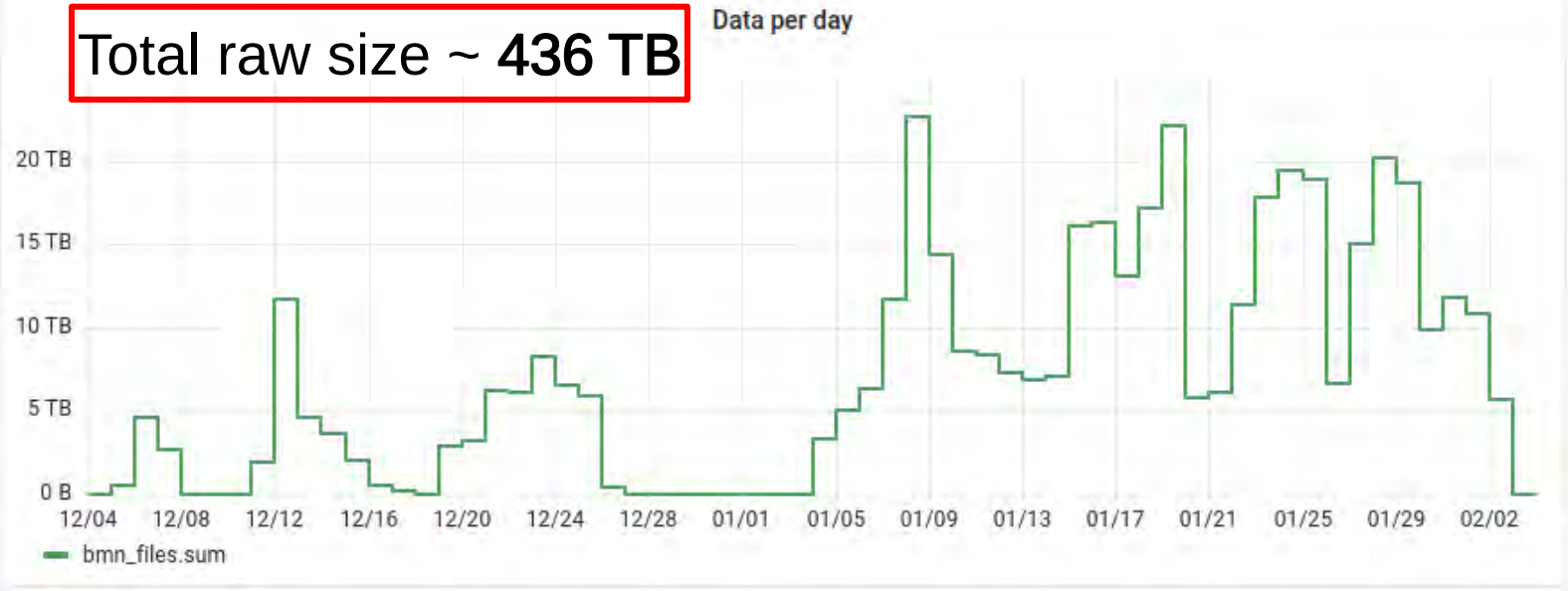
Konstantin Gertsenberger
LHEP

Igor Pelevanyuk
MLIT

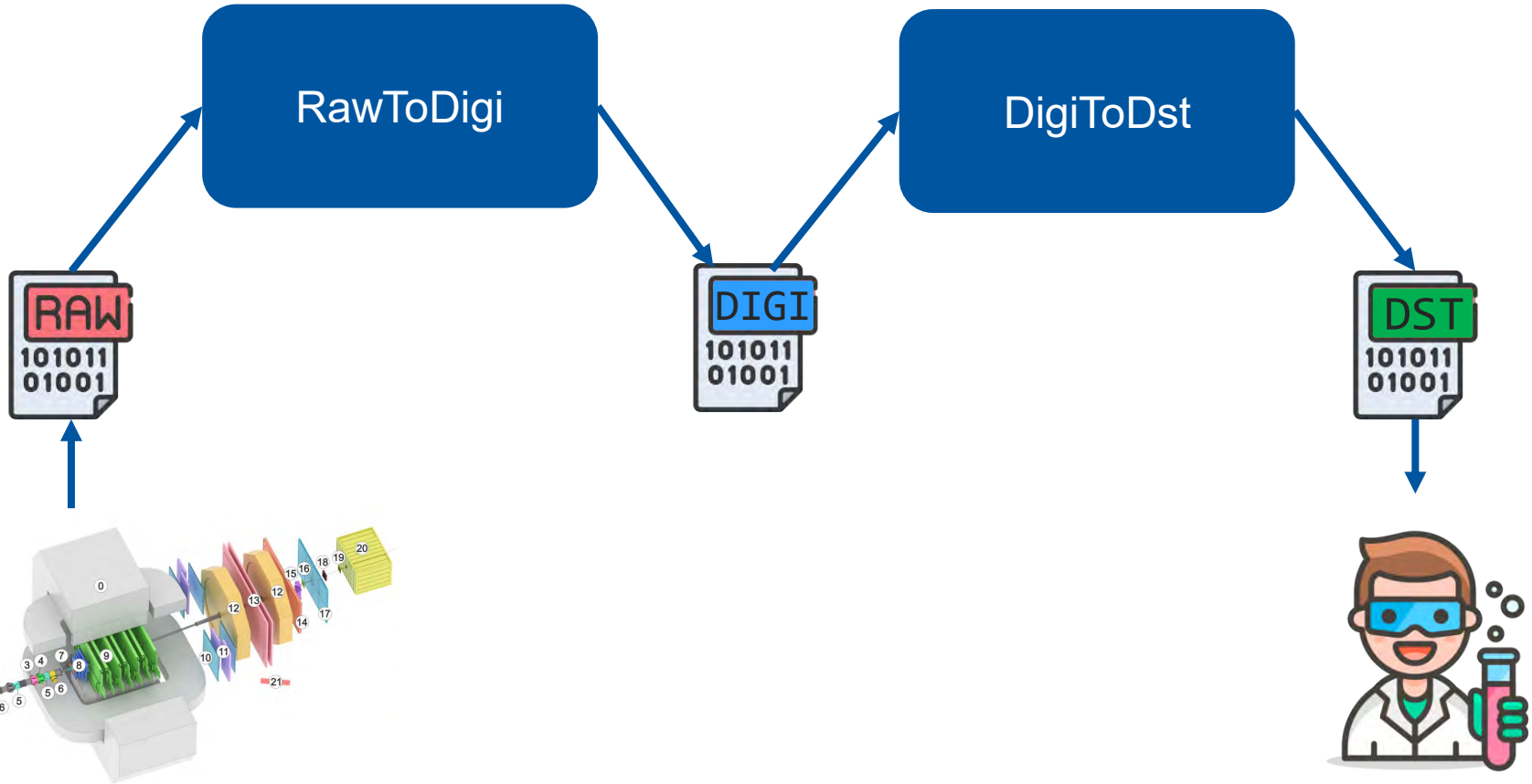
AYSS Conference 2024, 28 October - 01 November 2024

Run8 Data collection





Workflow of production



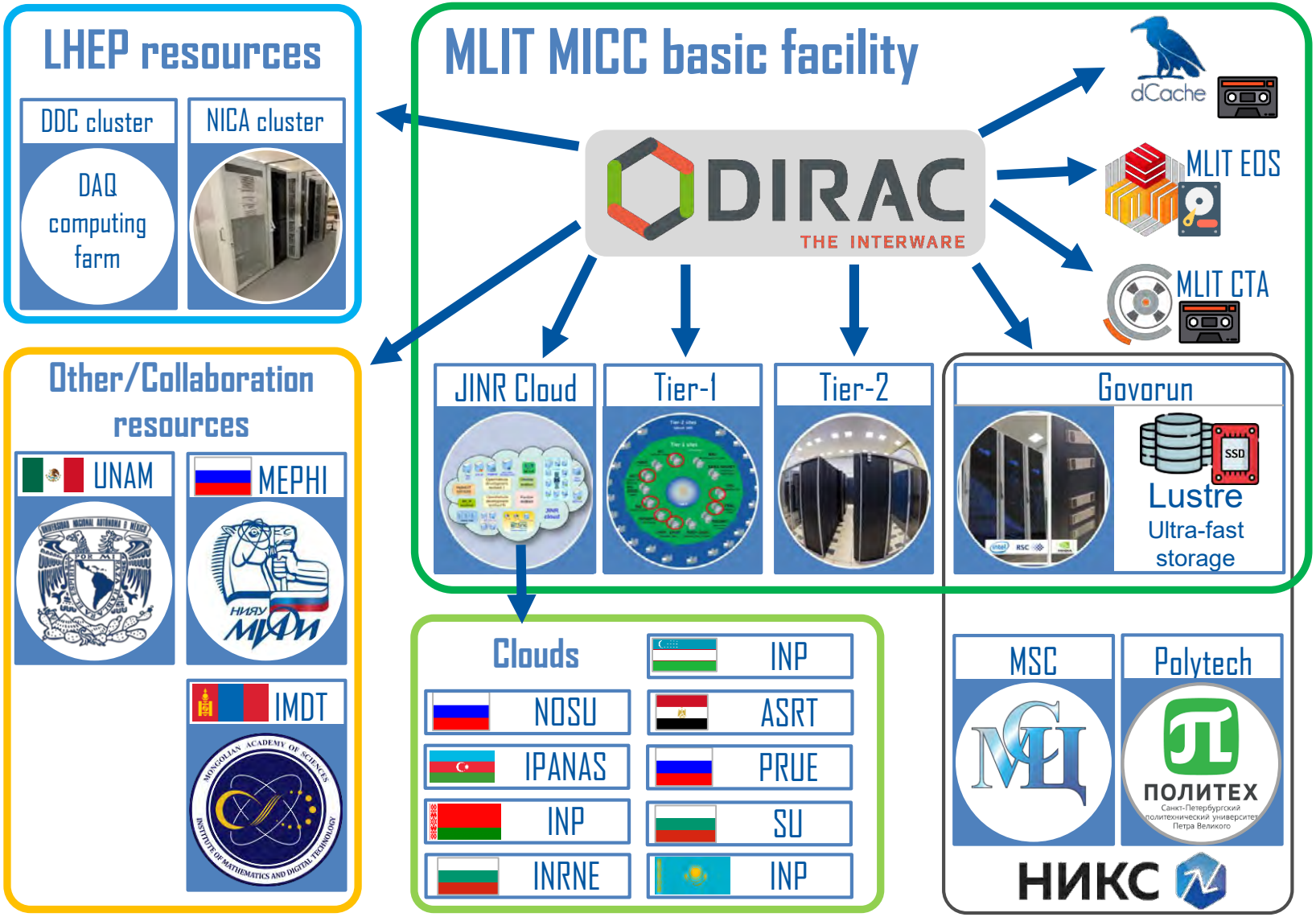
The main task:

Develop fast and repeatable way to consistently perform BM@N productions for all data in general, and for BM@N Run8 in particular

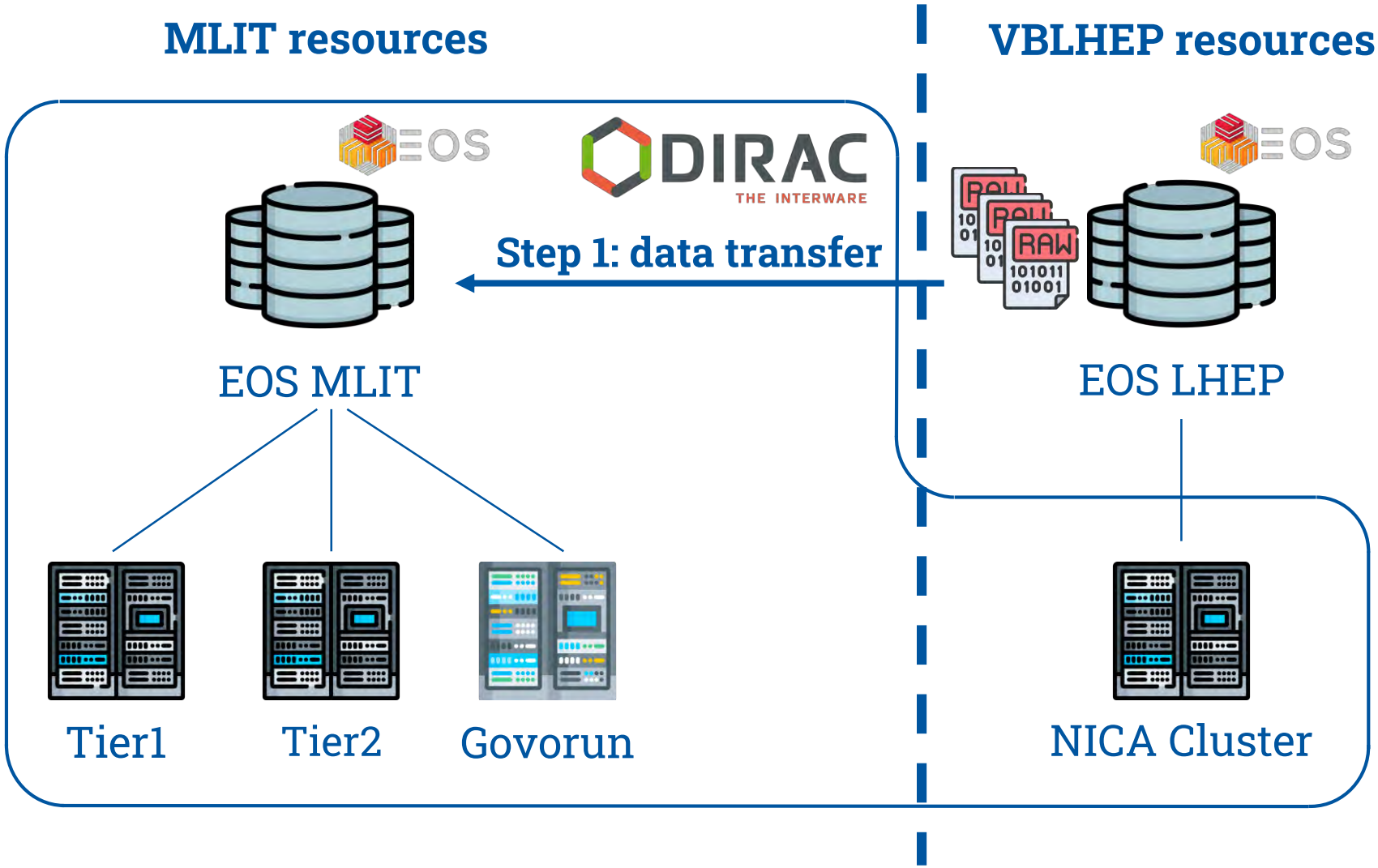
Side task:

Develop methods to record and use information about current productions for estimation of future productions

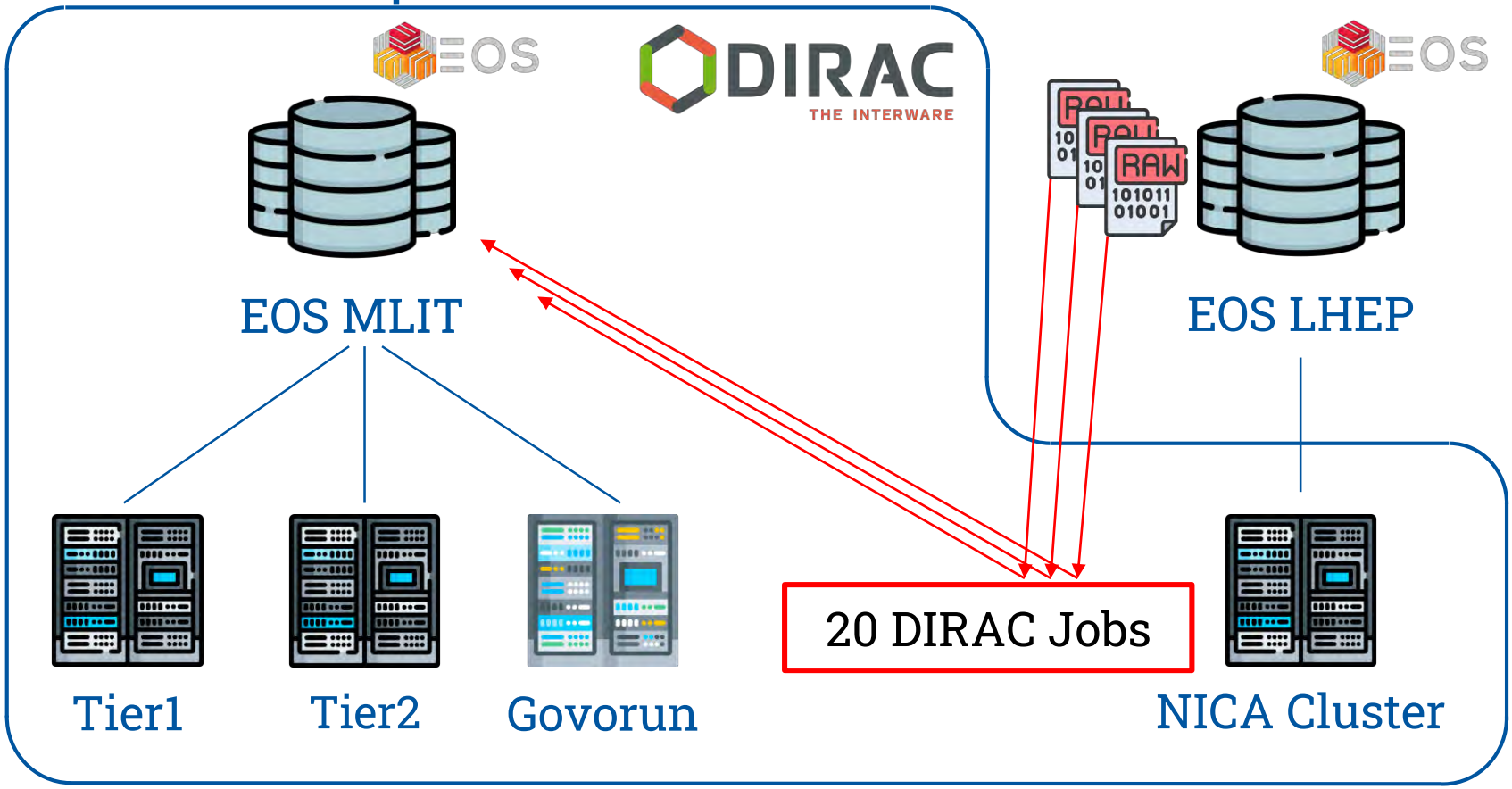
DIRAC in JINR



General scheme of resources



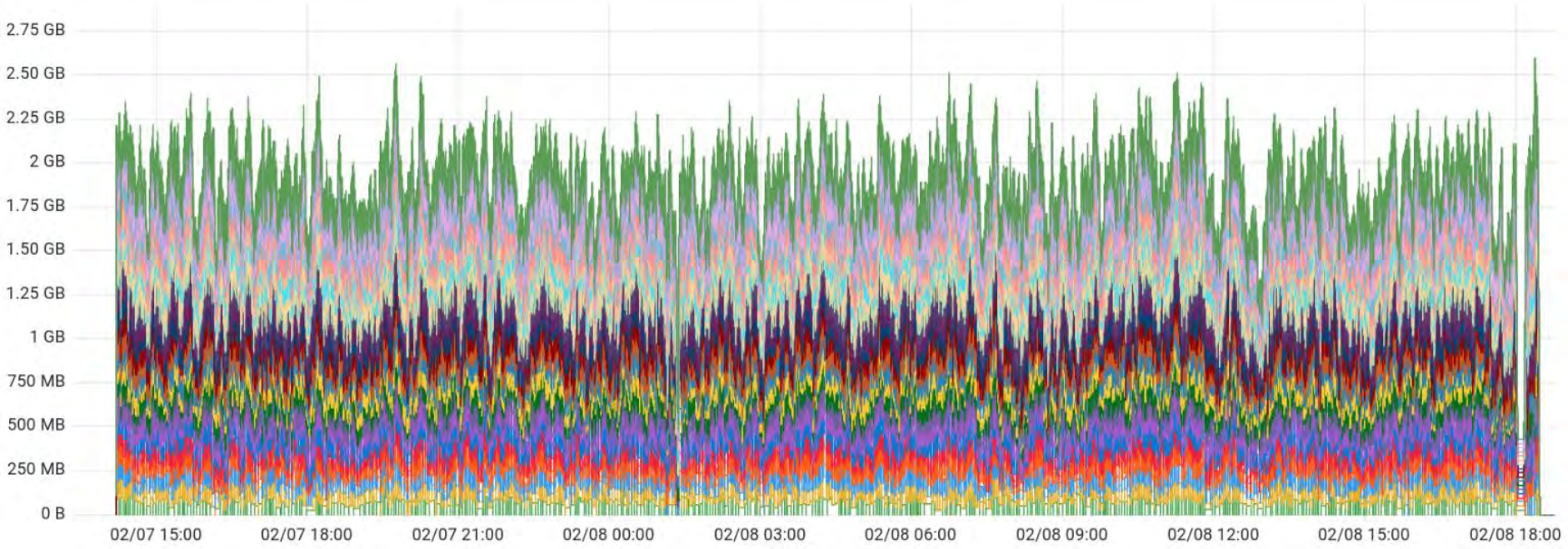
Step 1: Data transfer

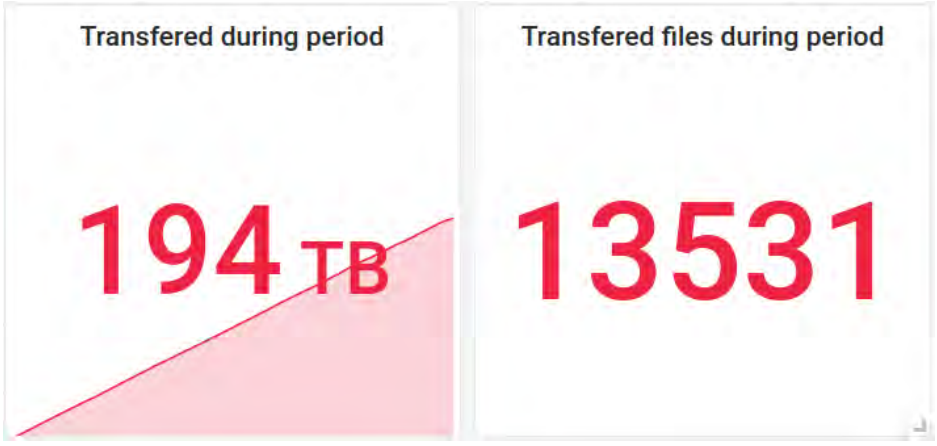


- Single stream of xrootd transfer can not exceed 100MB/s. **Transfer would take ~ 50 days.**
- NCX interface node can sustain not more than 10 streams(1GB/s total). And that would overload its network.

So, 20 independent DIRAC jobs were sent to NICA cluster to perform transfers with one stream each.

Step 1: Data transfer





Average transfer speed on 20 streams
1.92 GB/s

Total transfer duration:
2d 15h

Step 2: Estimate the load

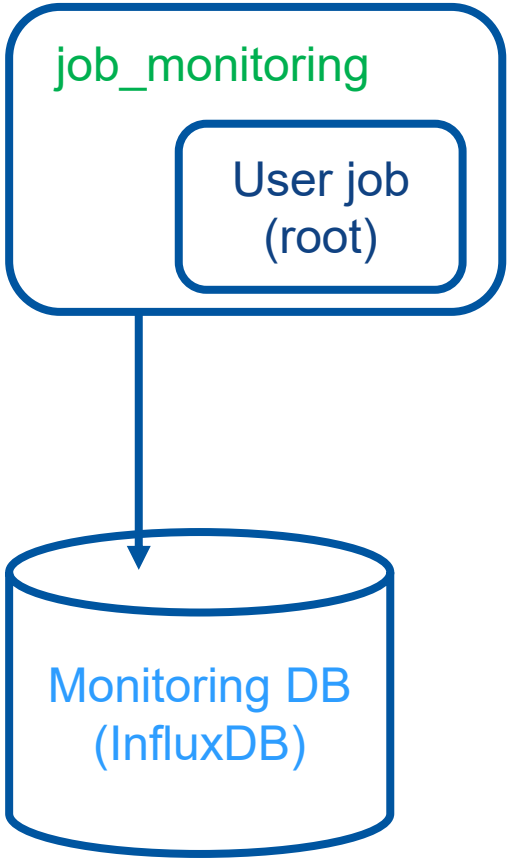
Size of files created during Run 8

Step 2: Raw2Digi job profiling

```
$ root macro.C(input)
```

Here is a standard process run

```
$ job_monitoring root macro.C(input)
```

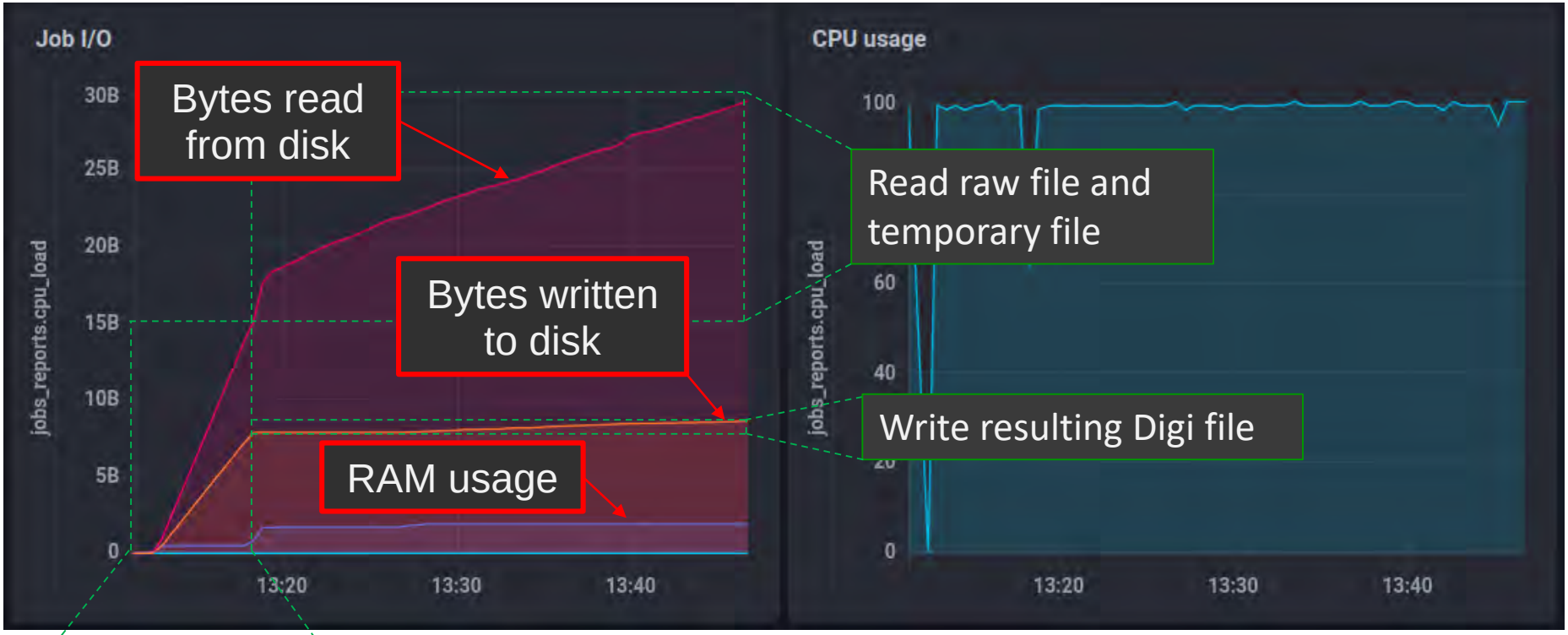


With the help of developed script we may record information about:

1. CPU used
2. RAM used
3. DISK read/write

Main issue was to record not only parameters of initial root process, but also its child processes.

Step 2: Raw2Digi job profiling



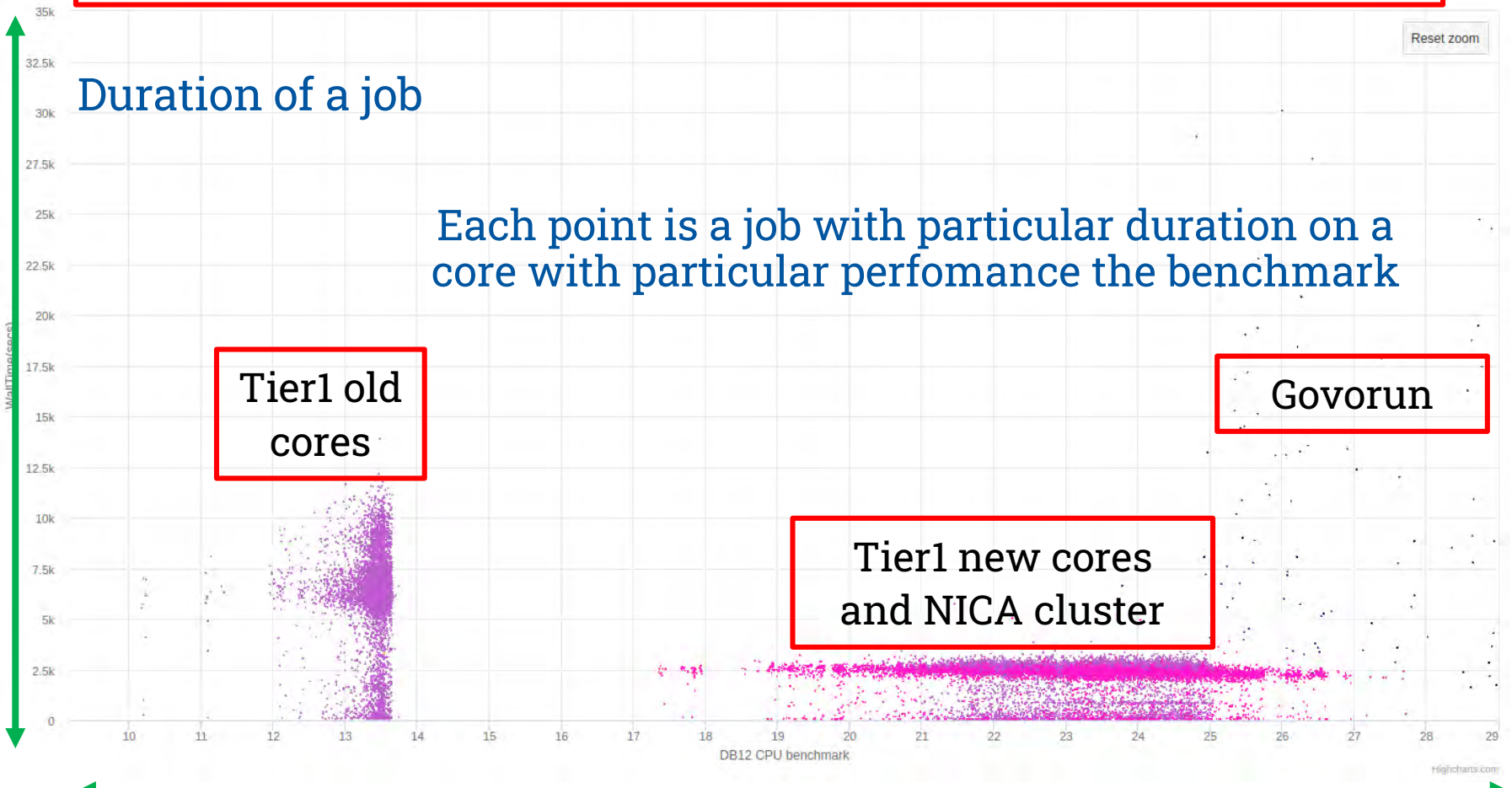
Initial read of 15GB raw file and creation of temporary 8 GB file

Disk usage
Temporary file: **+8 GB**
Result file: **800MB**
Total disk usage per 15 GB job: **25 GB**

RAM usage : **~2GB**

Step 3: Mass production Raw2Digi

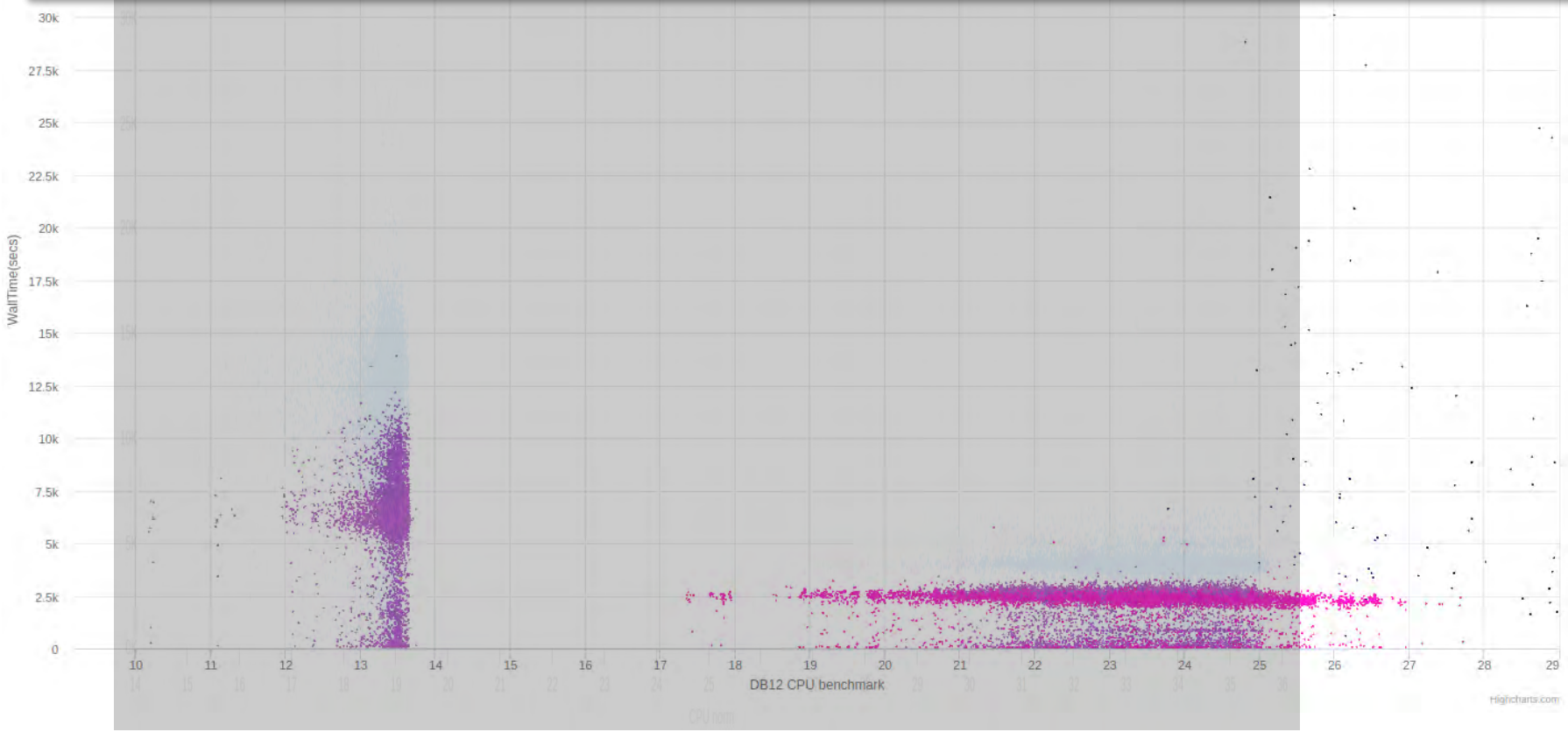
Total duration of Raw2Digi campaign – 35 hours



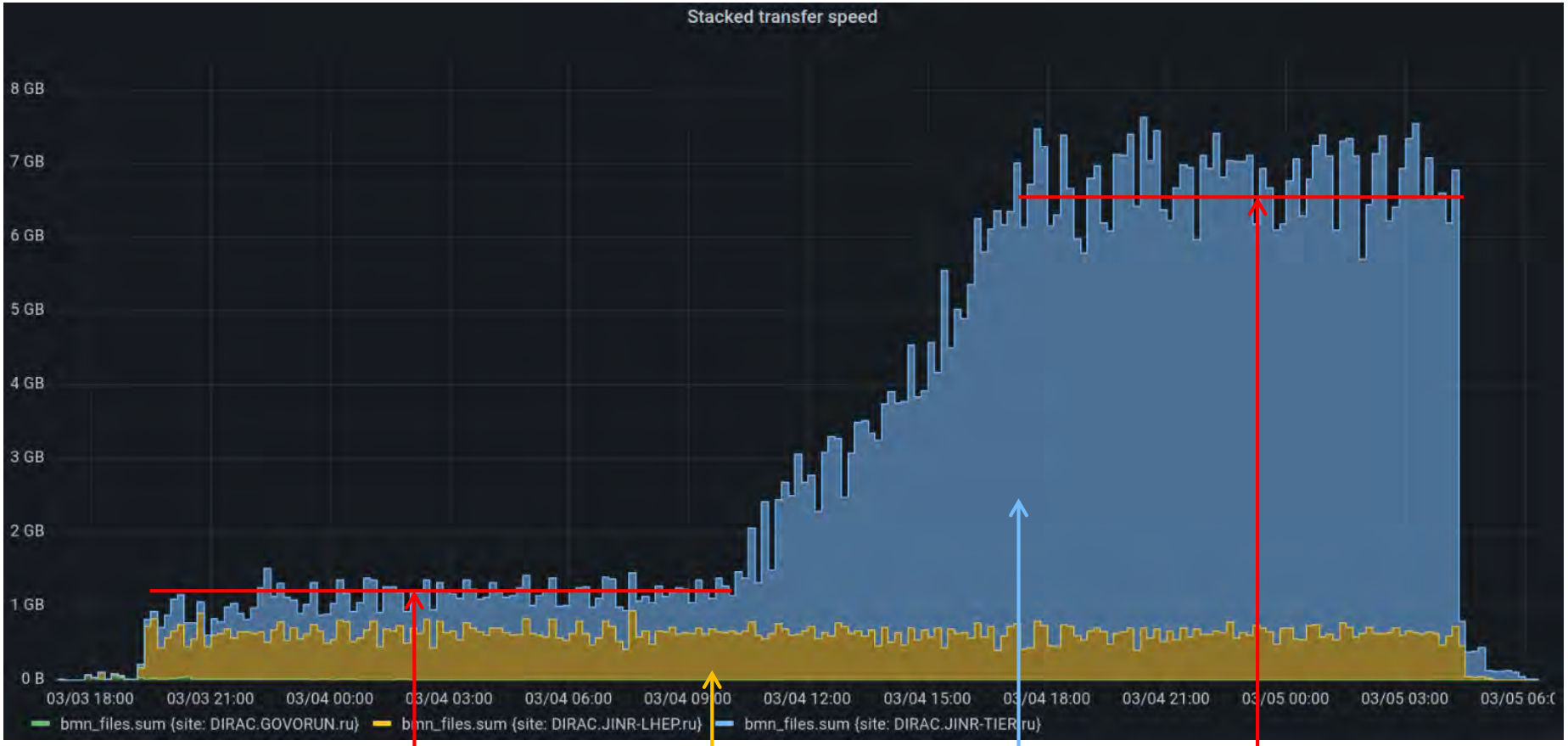
CPU core performance on benchmarks

Step 3: Mass production Raw2Digi

Average Raw2Digi calculation time increased by 60%
(from 35 hours to 56 hours)



Step 3: Network usage



300 jobs running
4 MB/s per job

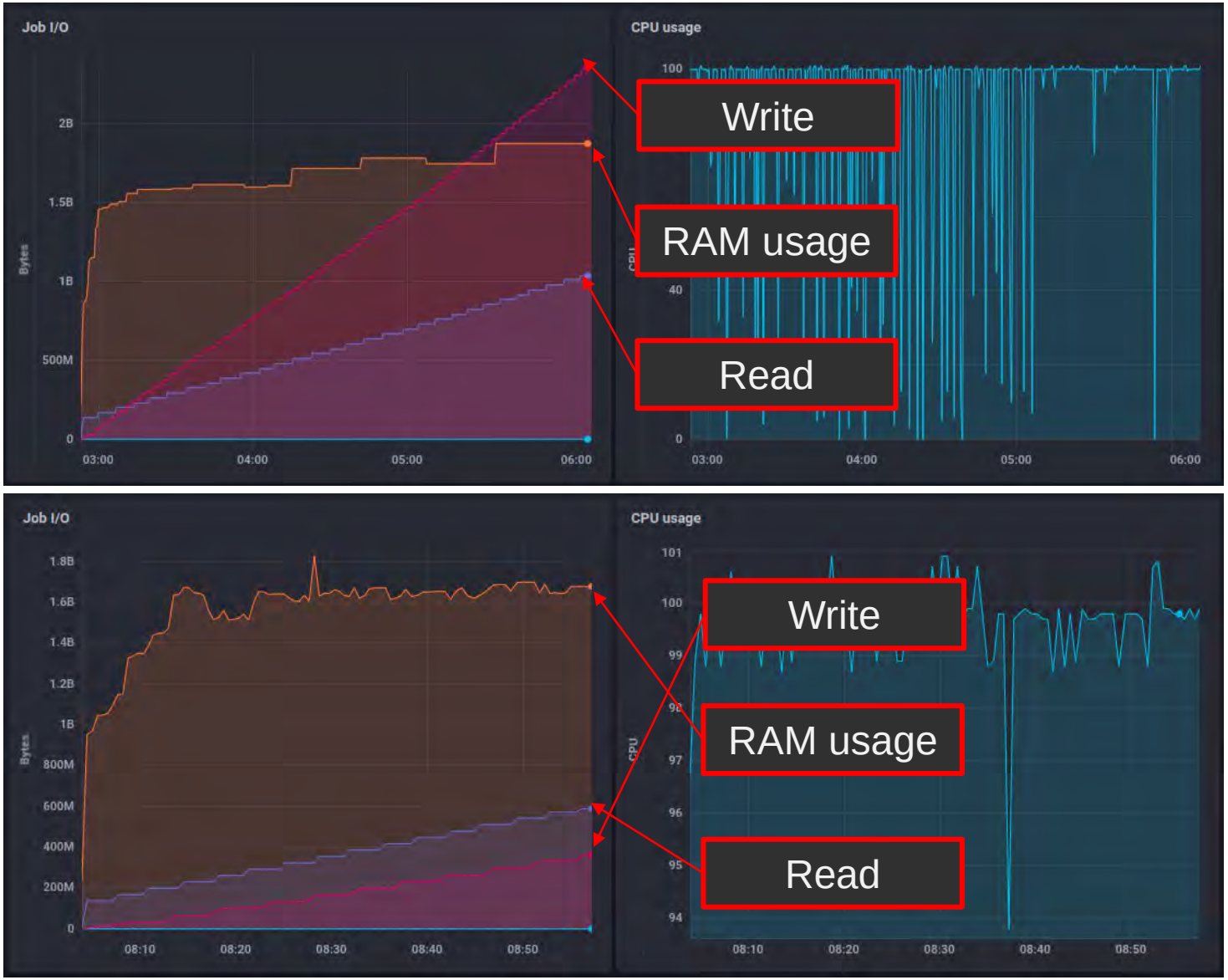
NICA Cluster

Tier1

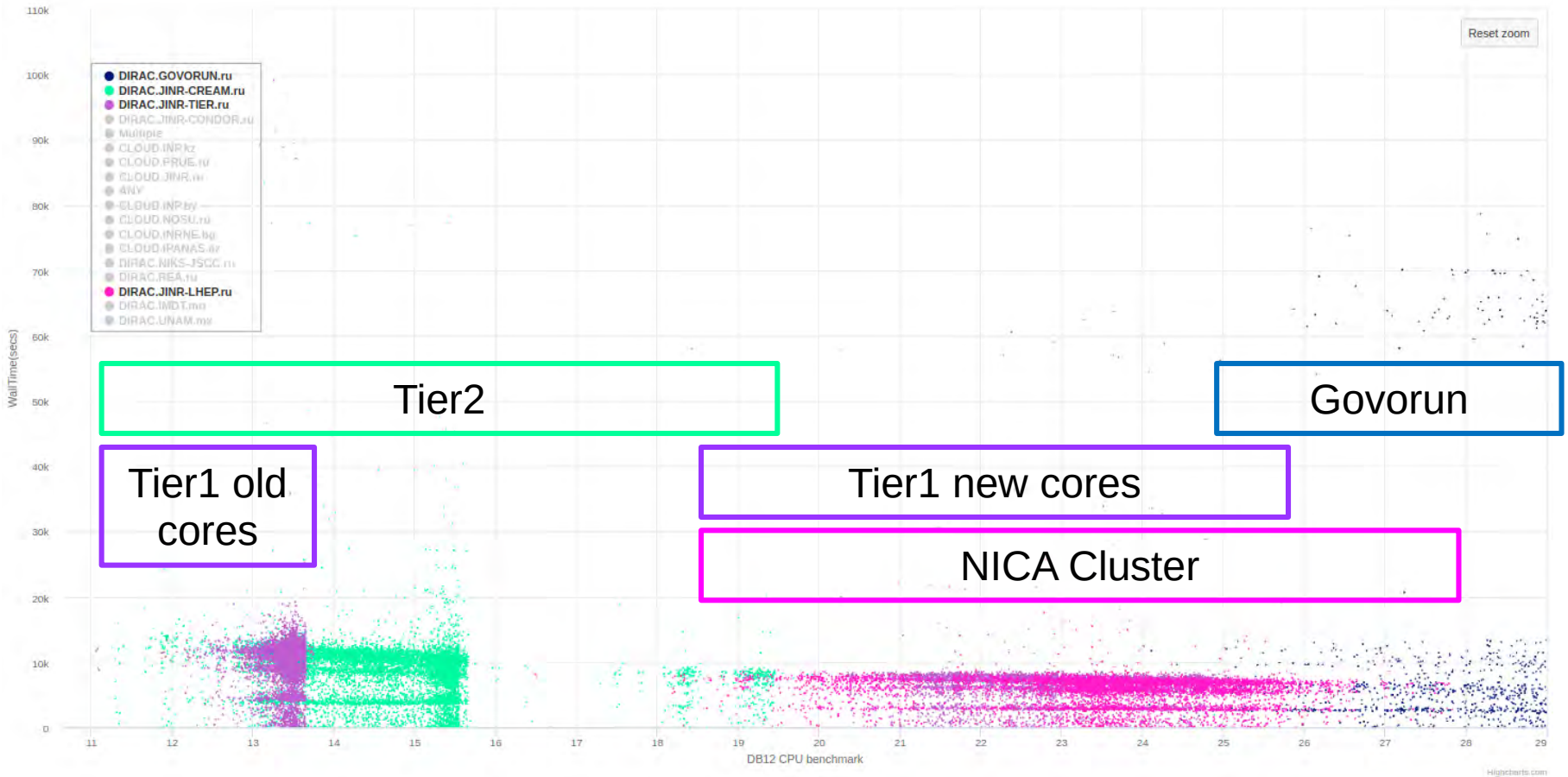
1580 jobs running
4.1 MB/s per job

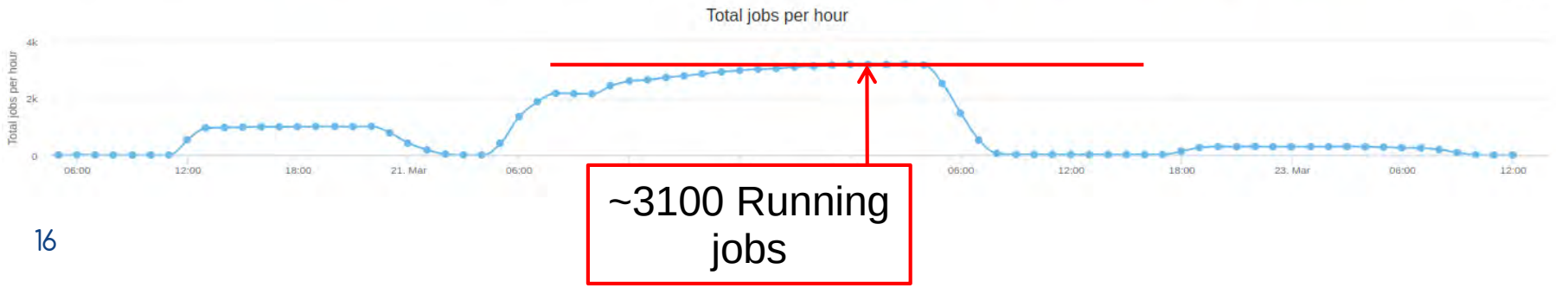
Maximal transfer speed (Read+Write) with EOS in MLIT – 7.5 GB/s

Step 4: Digi2Dst profiling



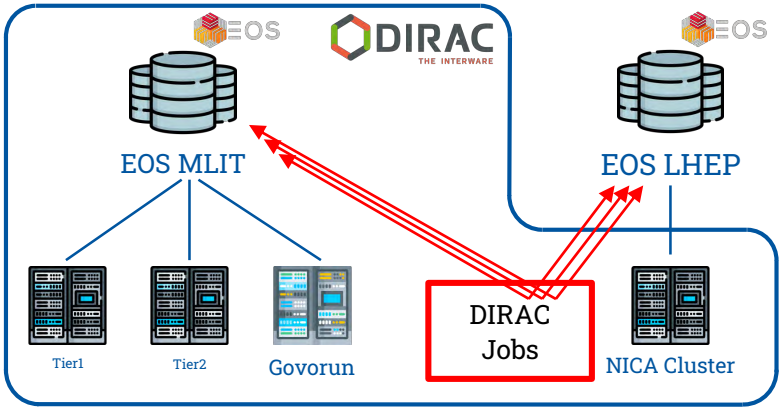
Step 5: Mass production Digi2Dst





What was achieved

Data transfer approach between VBLHEP and MLIT was developed and successfully used. It allows data transfer in both directions.

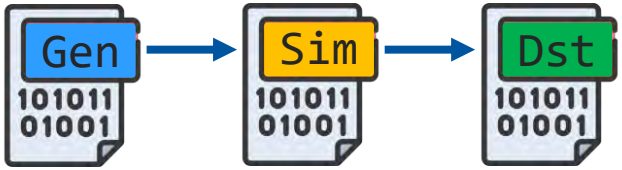


A set of standard workflows was prepared to perform:

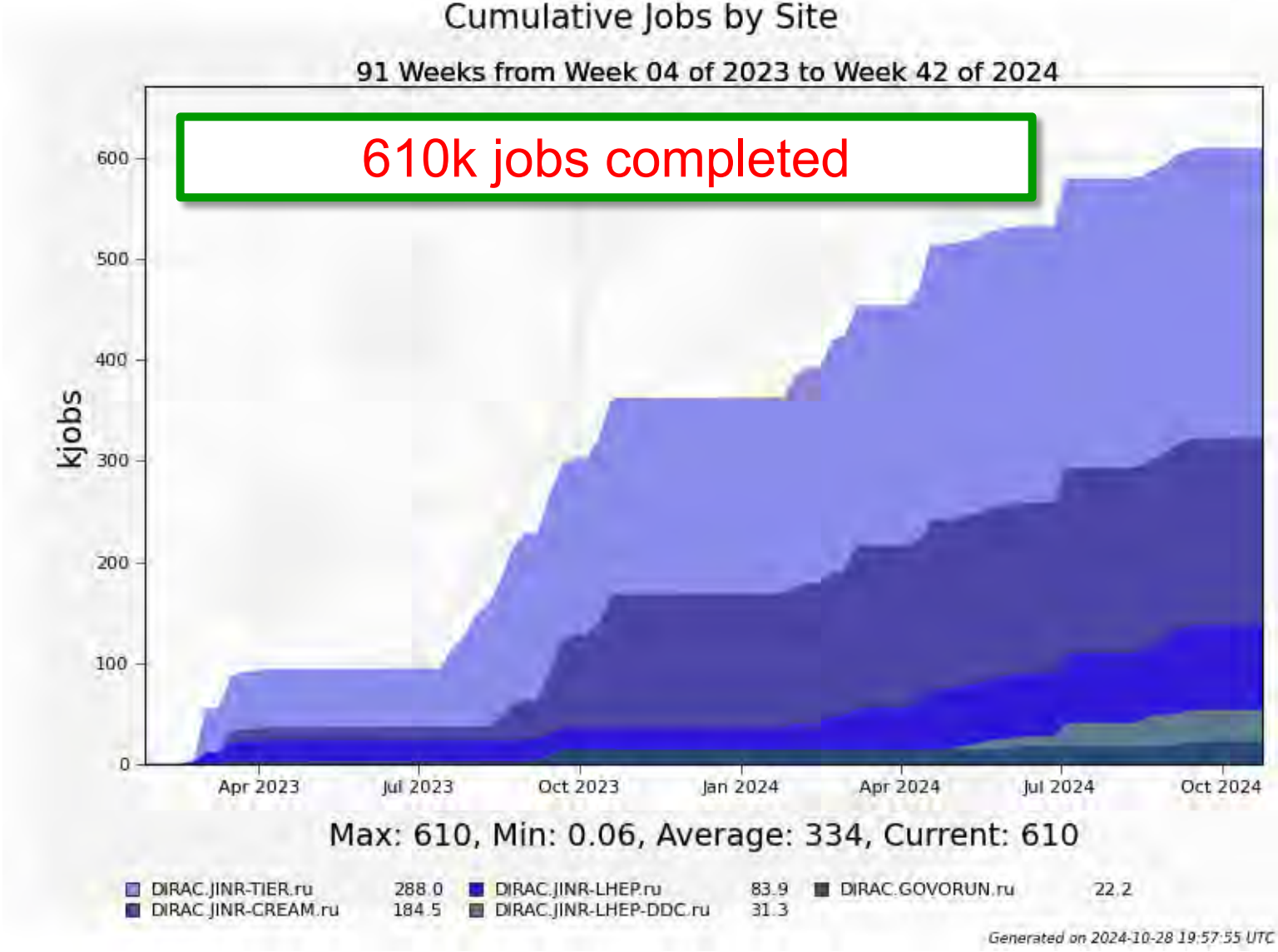
- **Raw to Digi** conversion
- **Digi to Dst** conversion

These was adopted for Monte-Carlo generation:

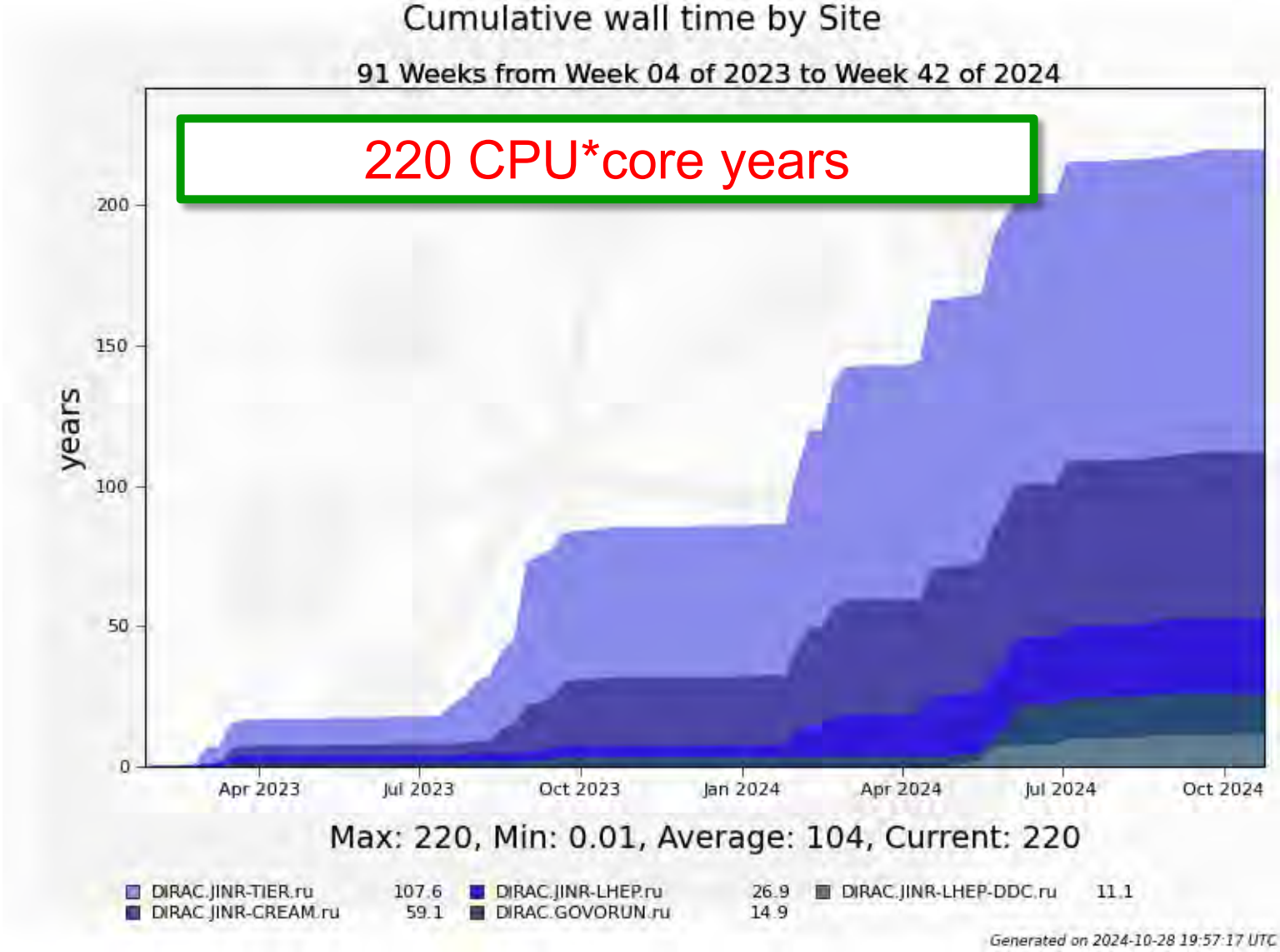
- **Gen to Sim**
- **Sim to Dst**
- **Gen to Dst**



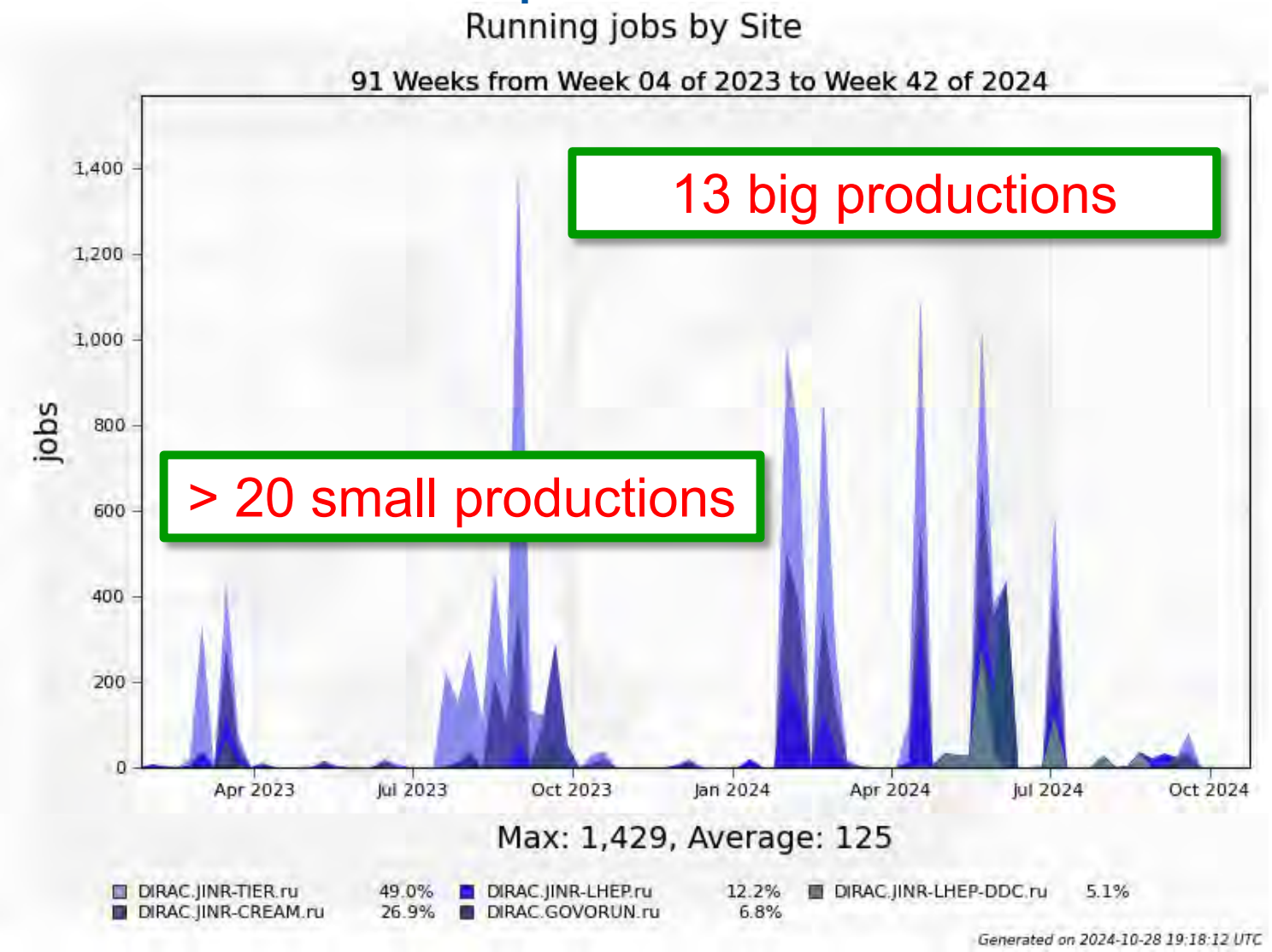
Completed jobs



Consumed resources



BM@N productions



Results

Fast and repeatable way to consistently perform B@MN productions for all data was proposed!

It was successfully applied many times for BM@N Run8 data.

A set of methods was developed and applied to record the information about running productions.

It proved to be useful not only for estimation of future productions, but also for evaluation of current workloads and their comparison with previous.

- For the first time JINR computing infrastructure united by DIRAC was used for raw data reconstruction not in test mode but in production.
- For the last 1.5 years BM@N used considerable amount of resources applying developed approaches. Up to now ~220 CPU core years has been consumed, 610k jobs completed.

Acknowledgments

The whole **BM@N** collaboration

Responsible for resources:
Tier-1,Tier-2, EOS: Valery Mitsyn
Tape library: Vladimir Trofimov
Govorun: Dmitry Podgainy, Oksana Smirnova, Dmitry Belyakov, Aleksandr Kokorev, Maxim Zuev
NICA cluster: Ivan Slepov
Network: Andrey Dolbilov
CVMFS: Nikita Balashov


JINR
MLiT

Individual CPU core performance study

- Centralized job management gives possibility for centralized and unified performance study of different computing resources.
- Before running user jobs DIRAC Pilots execute benchmark for CPU core they are running on.
- Benchmark is DiracBenchmark2012 or DB12. It evaluate just CPU core performance. Disk I/O, RAM speed, Network, CPU caches and other highly important aspects of performance are **neglected by DB12.**

DB12 benchmark study

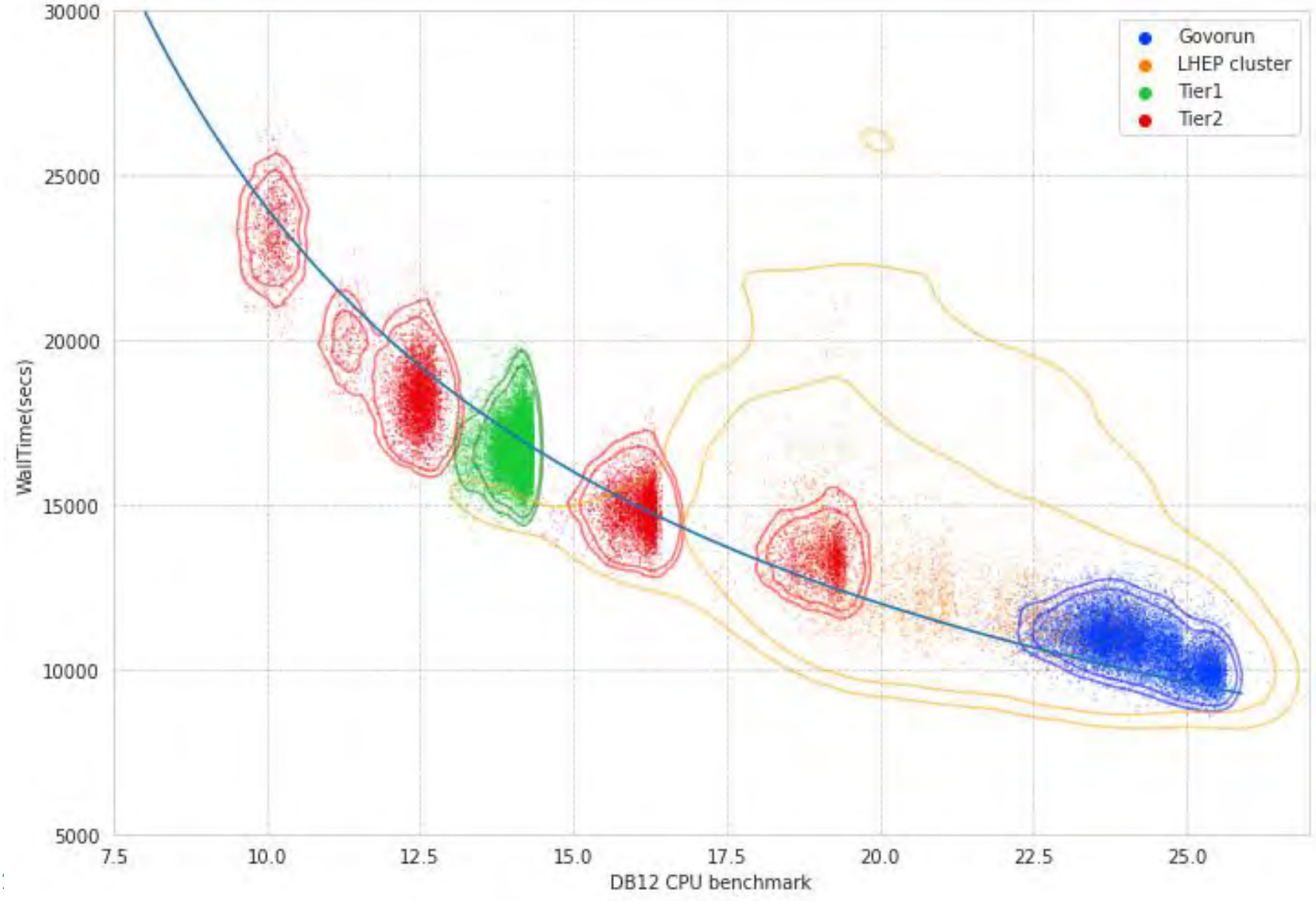
Piece of road from point A to B	←	A Monte-Carlo task
Speed of the car	←	Performance of the computer
Time to complete	←	Time to complete

$$Time = \frac{Amount\ of\ work}{Speed\ of\ computer}$$

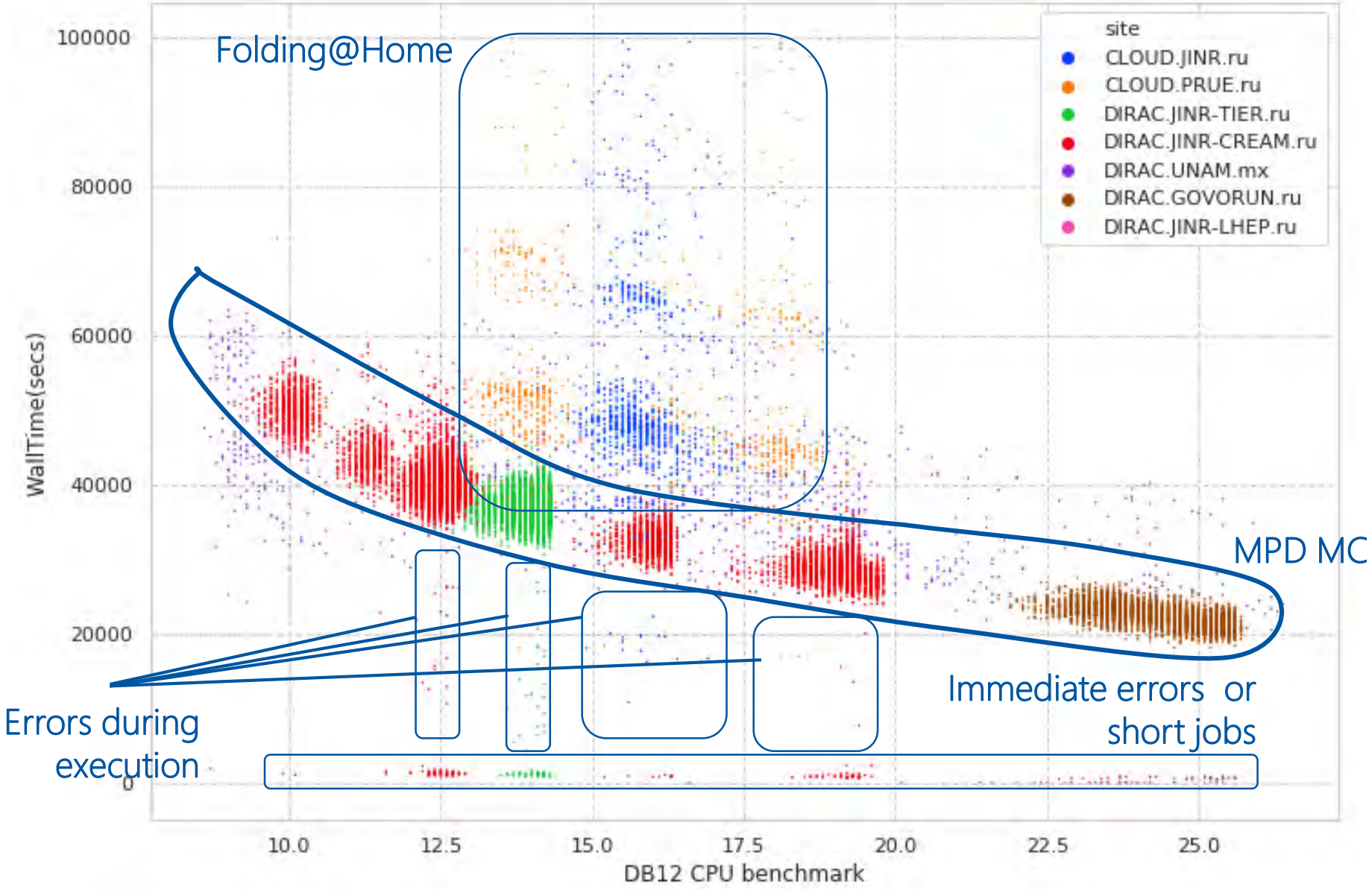
DB12 gives results like: 10(old slow core), 17 (standard server core), 27 (high performance core)

What if we build a plot, where X is DB12 result, Y is time in seconds. Then, every point on the plot represent one job.
It would be mostly useless if all jobs were unique and different.
But, in the real life there are usually many similar jobs.

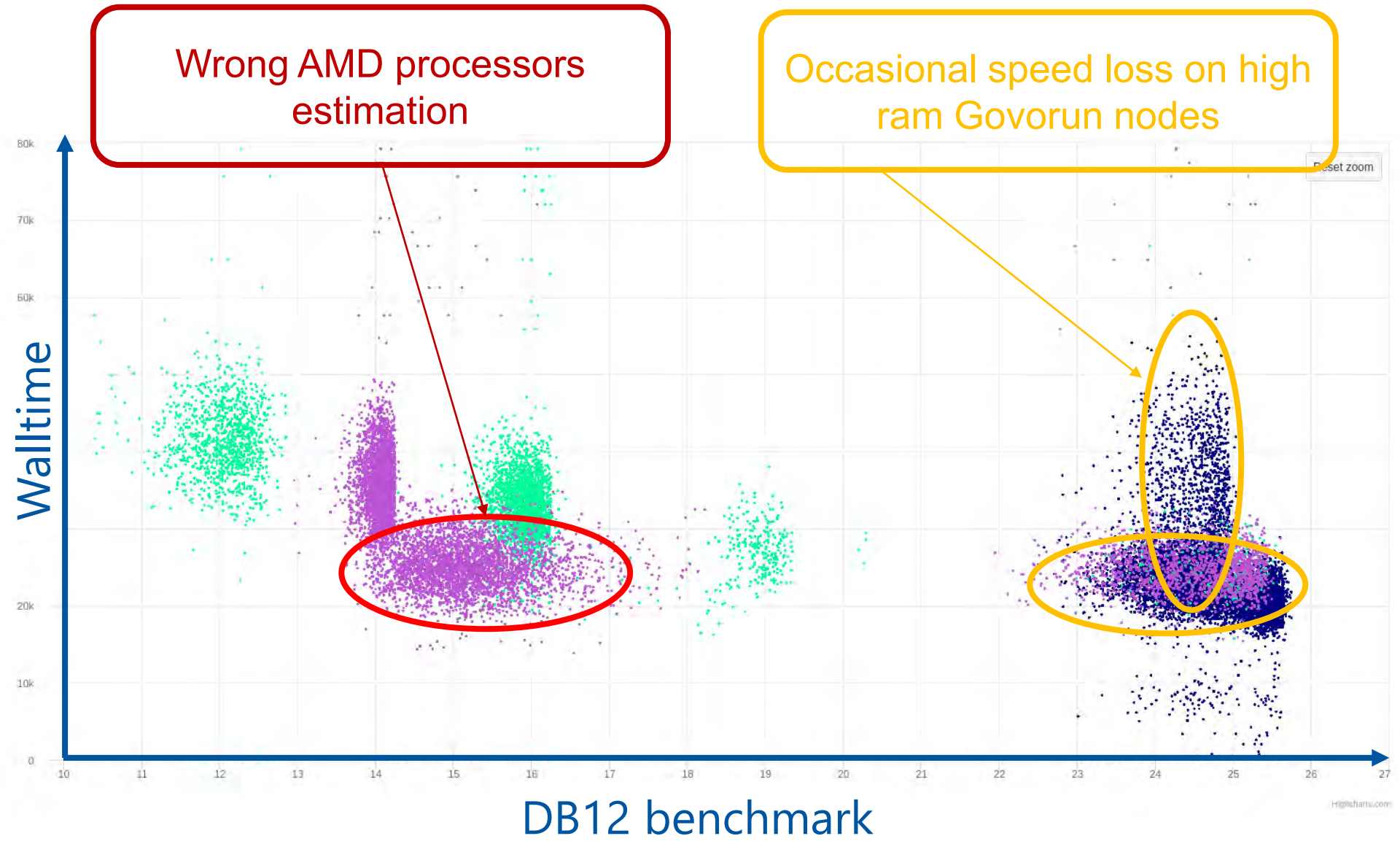
Performance analysis



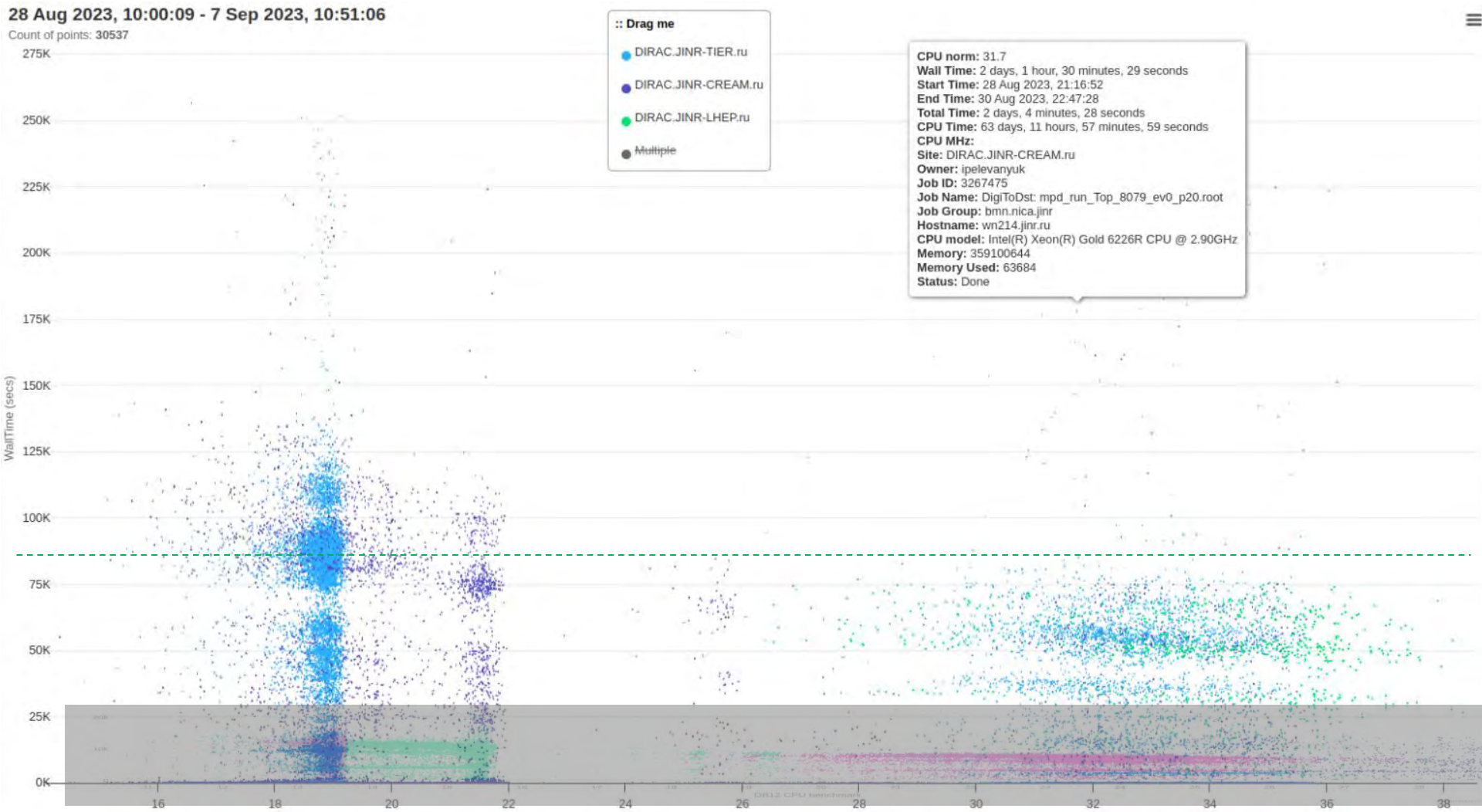
Performance analysis



Discoveries



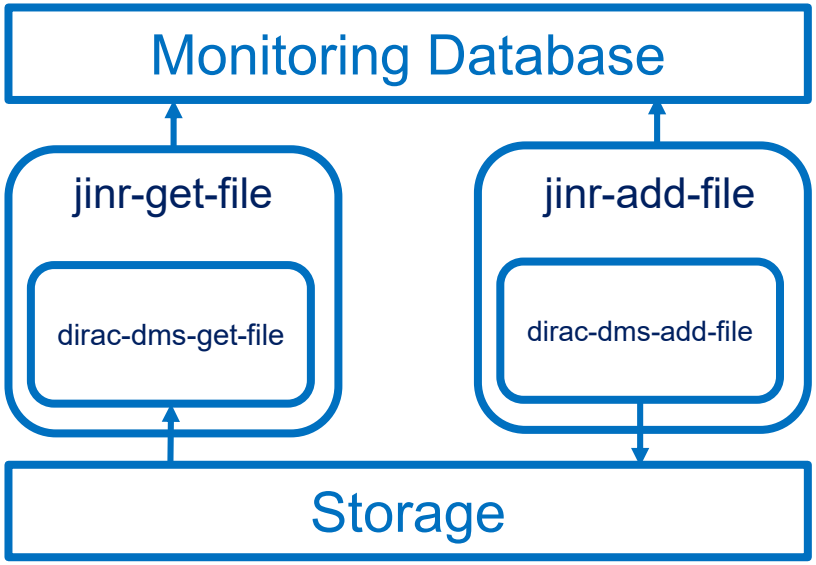
Step 5: New VF Digi2Dst

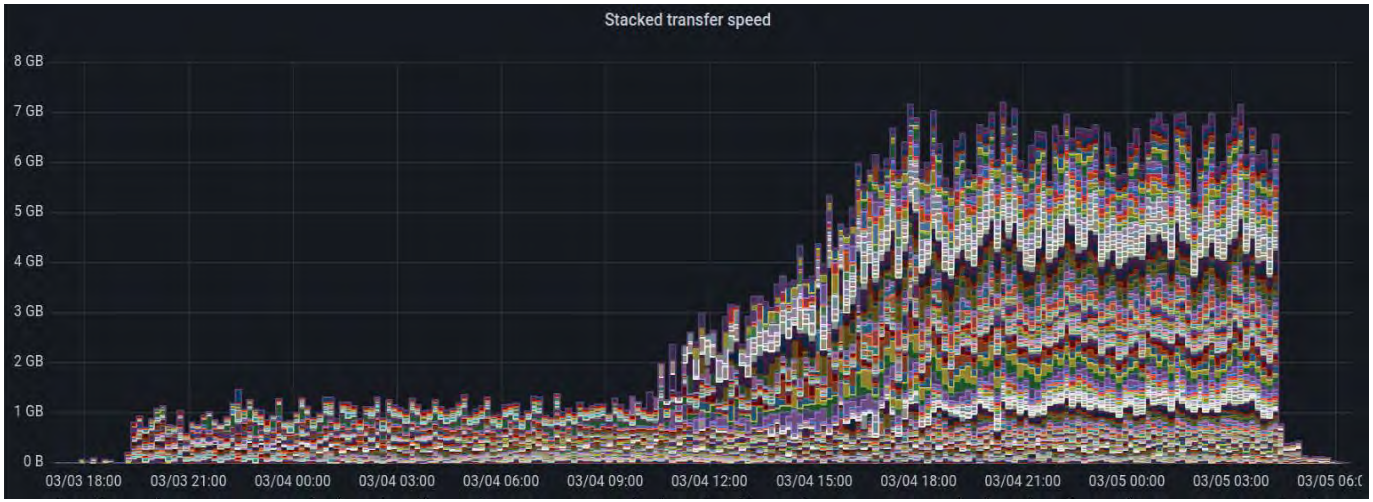


CPU core performance

Total CPU performance

Data transfers monitoring



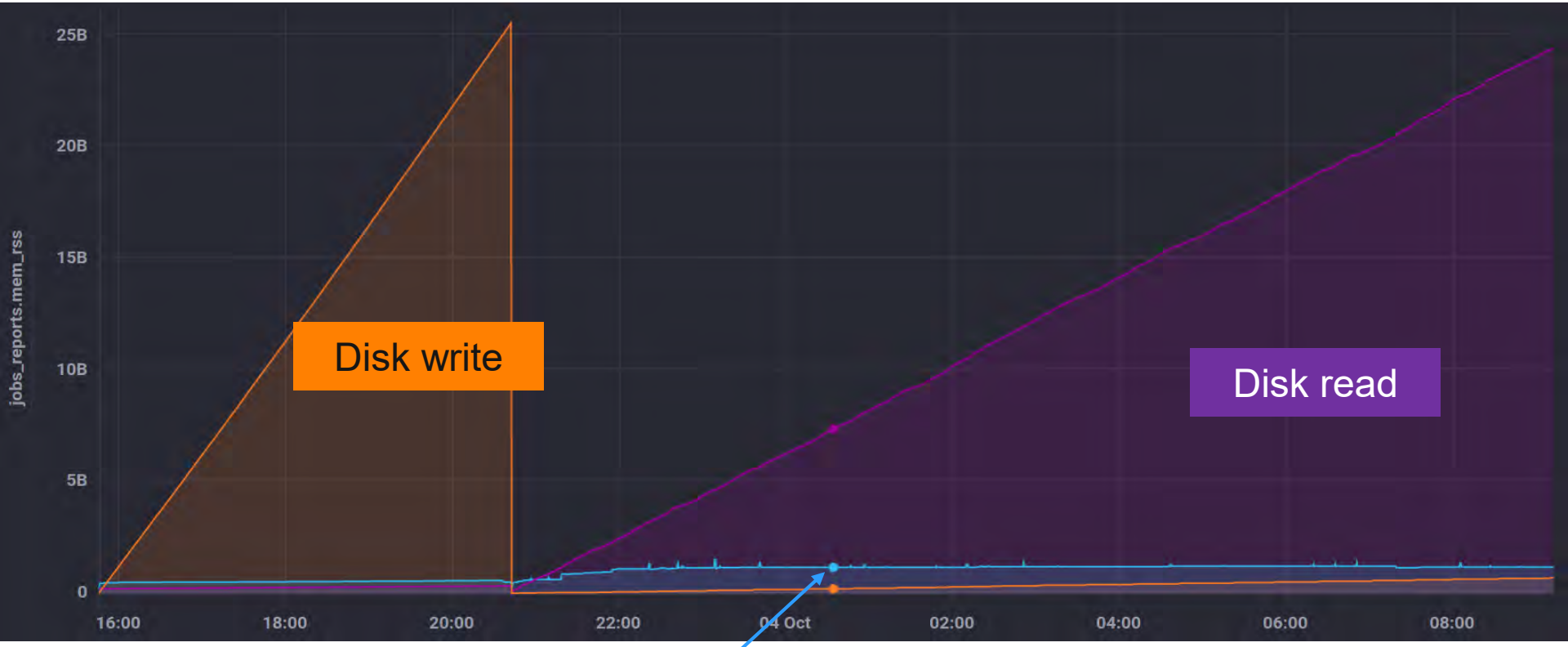


Step 2: Raw2Digi job profiling

Once Raw2Digi had strange 5m idle period.
Reason is a request to a database which failed by timeout.

User job monitoring

GenToDst job on Govorun



Detailed articles

1. Gergel, V., V. Korenkov, I. Pelevanyuk, M. Sapunov, A. Tsaregorodtsev, and P. Zrelov. 2017. **Hybrid Distributed Computing Service Based on the DIRAC Interware**.
2. Korenkov, V., Pelevanyuk, I. & Tsaregorodtsev, A. 2019, "**Dirac system as a mediator between hybrid resources and data intensive domains**", CEUR Workshop Proceedings, pp. 73.
3. Balashov, N.A., Kuchumov, R.I., Kutovskiy, N.A., Pelevanyuk, I.S., Petrunin, V.N. & Tsaregorodtsev, A.Y. 2019, "**Cloud integration within the DIRAC Interware**", CEUR Workshop Proceedings, pp. 256.
4. Korenkov, V., Pelevanyuk, I. & Tsaregorodtsev, A. 2020, **Integration of the JINR hybrid computing resources with the DIRAC interware for data intensive applications**.
5. Kutovskiy, N., Mitsyn, V., Moshkin, A., Pelevanyuk, I., Podgayny, D., Rogachevsky, O., Shchinov, B., Trofimov, V. & Tsaregorodtsev, A. 2021, "**Integration of Distributed Heterogeneous Computing Resources for the MPD Experiment with DIRAC Interware**", Physics of Particles and Nuclei, vol. 52, no. 4, pp. 835-841.
6. Pelevanyuk, I., "**Performance evaluation of computing resources with DIRAC interware**", AIP Conference Proceedings 2377, 040006 (2021)