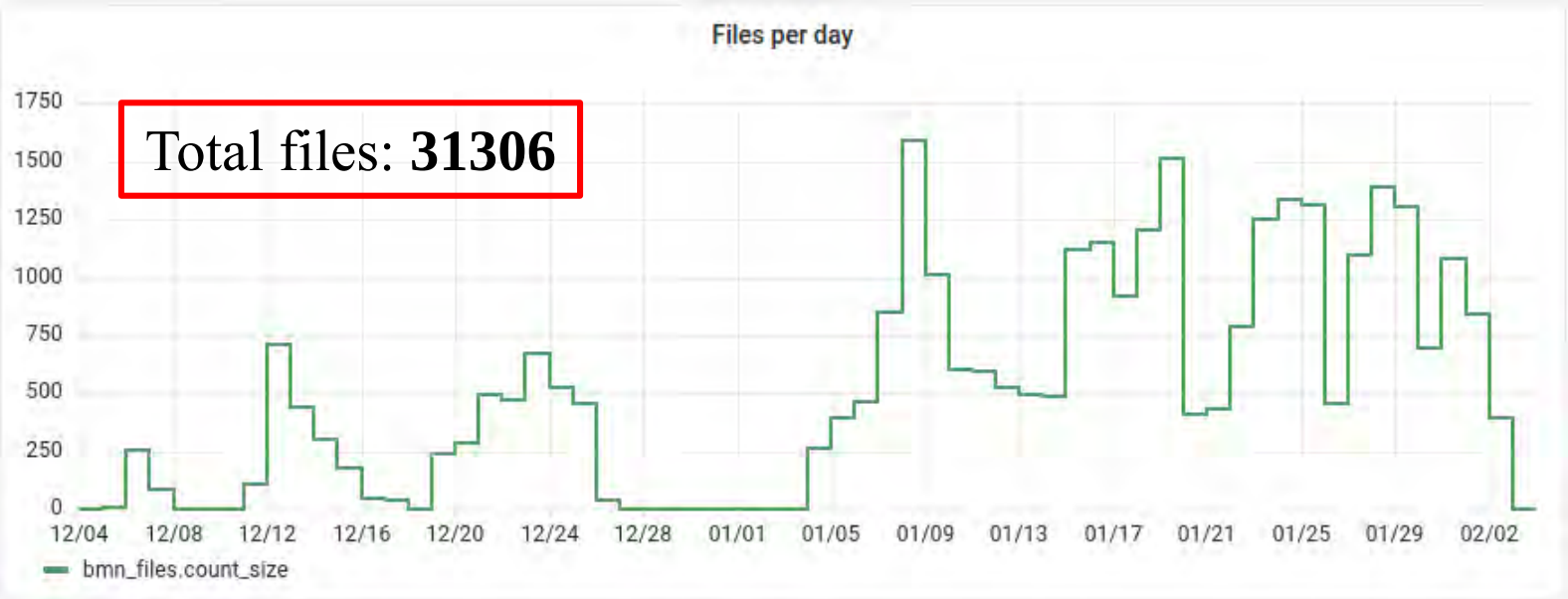
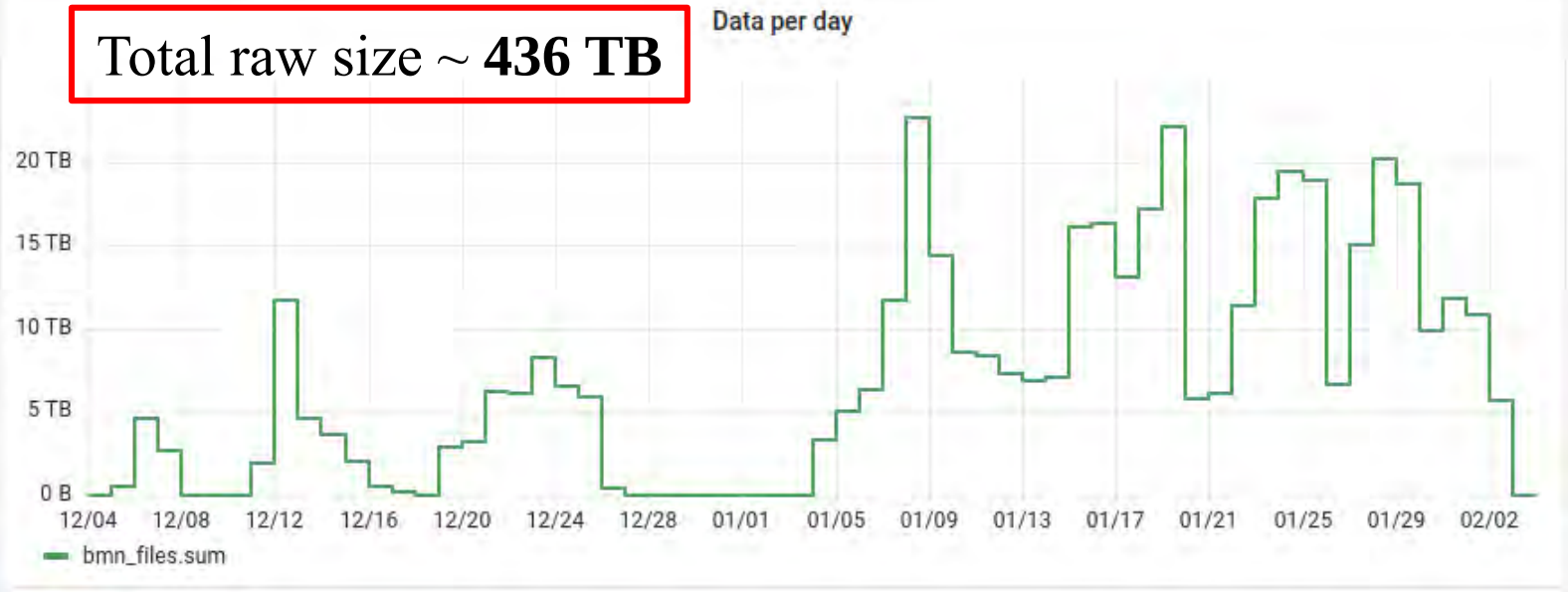
BM@N Run 8 raw data reconstruction on distributed infrastructure with DIRAC

Konstantin Gertsenberger , Igor Pelevanyuk
LHEP MLIT

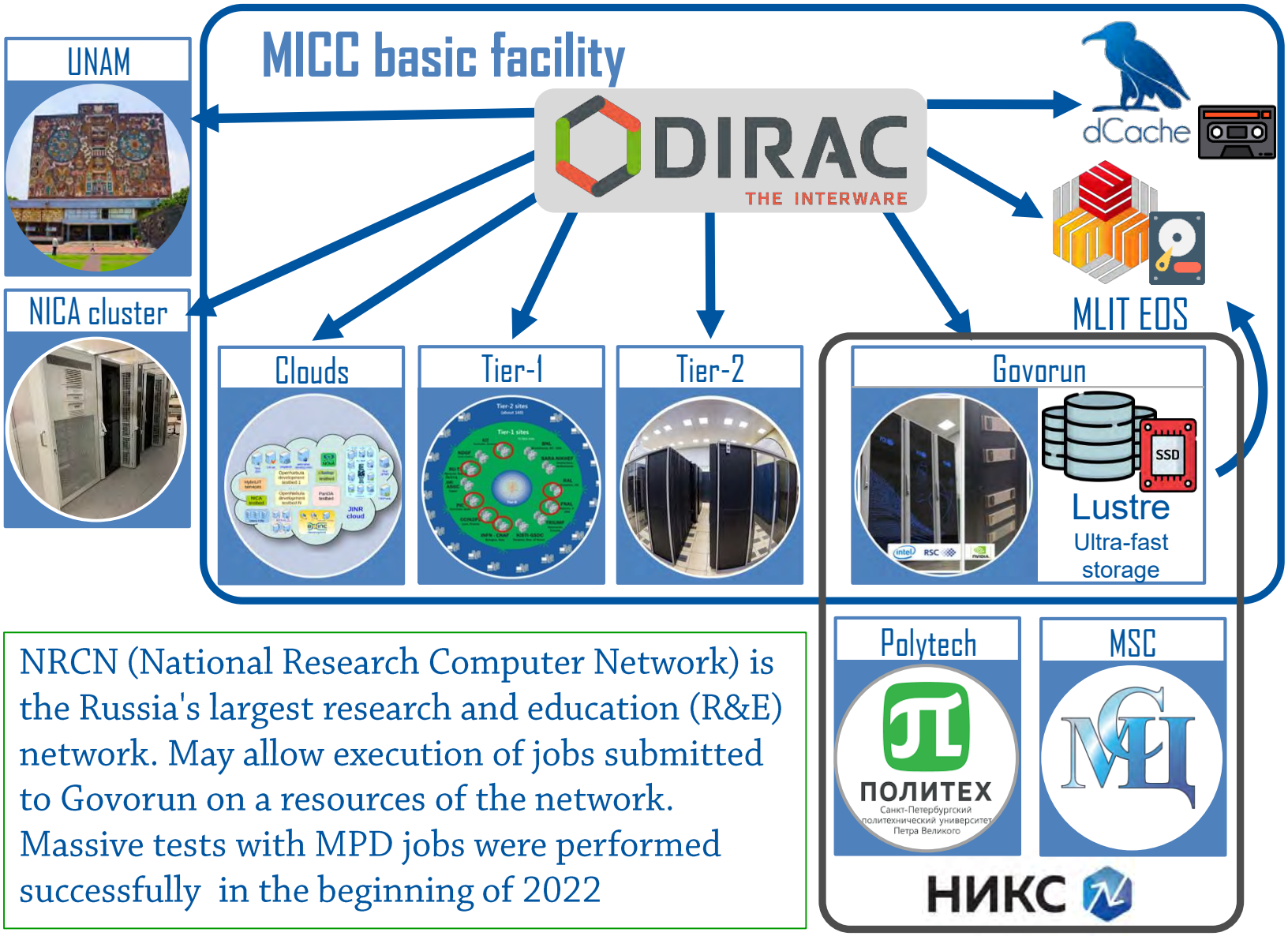
10th BM@N Collaboration Meeting, 14-19 May 2023,

Run8 Data collection



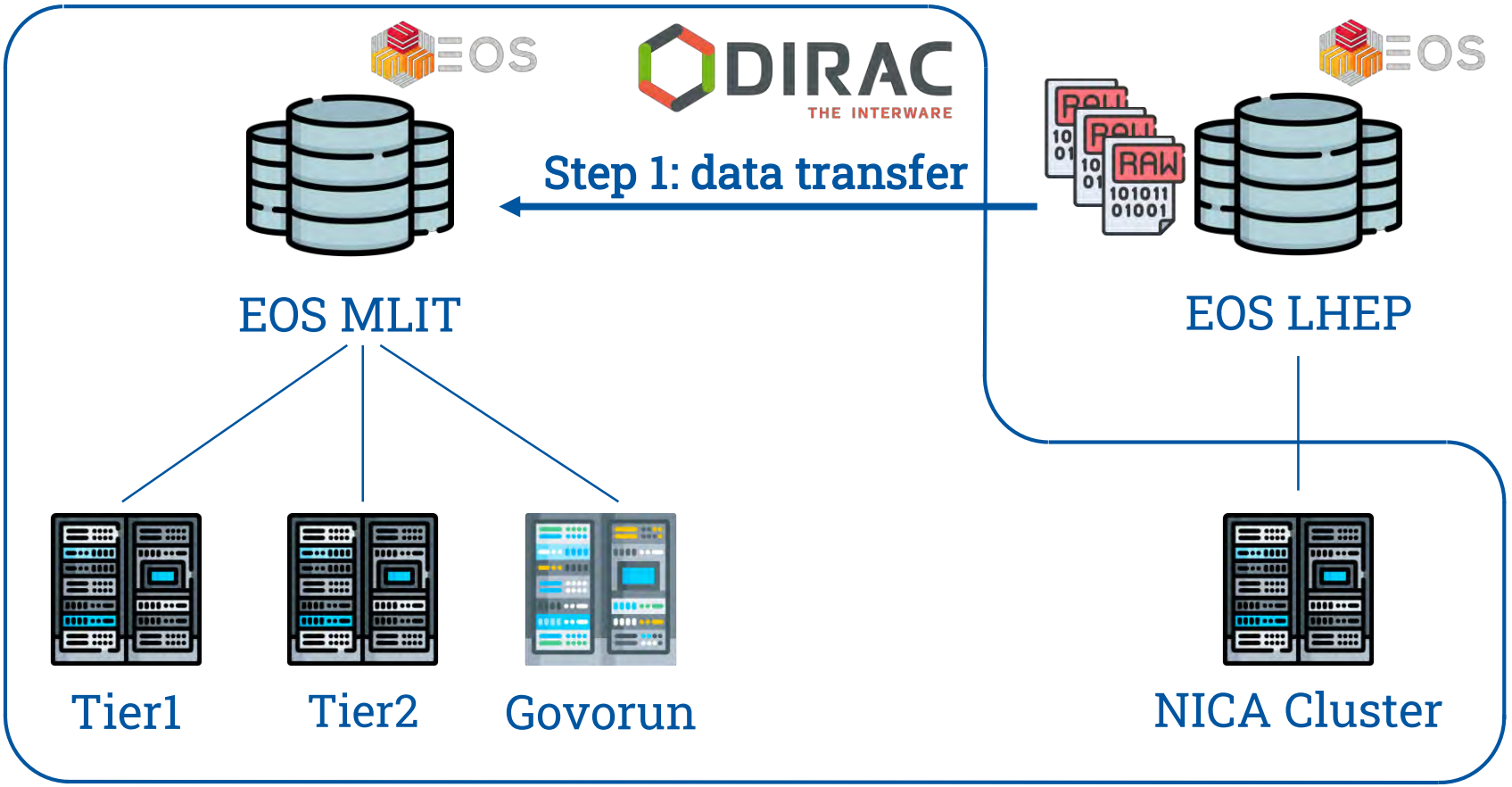


DIRAC in JINR

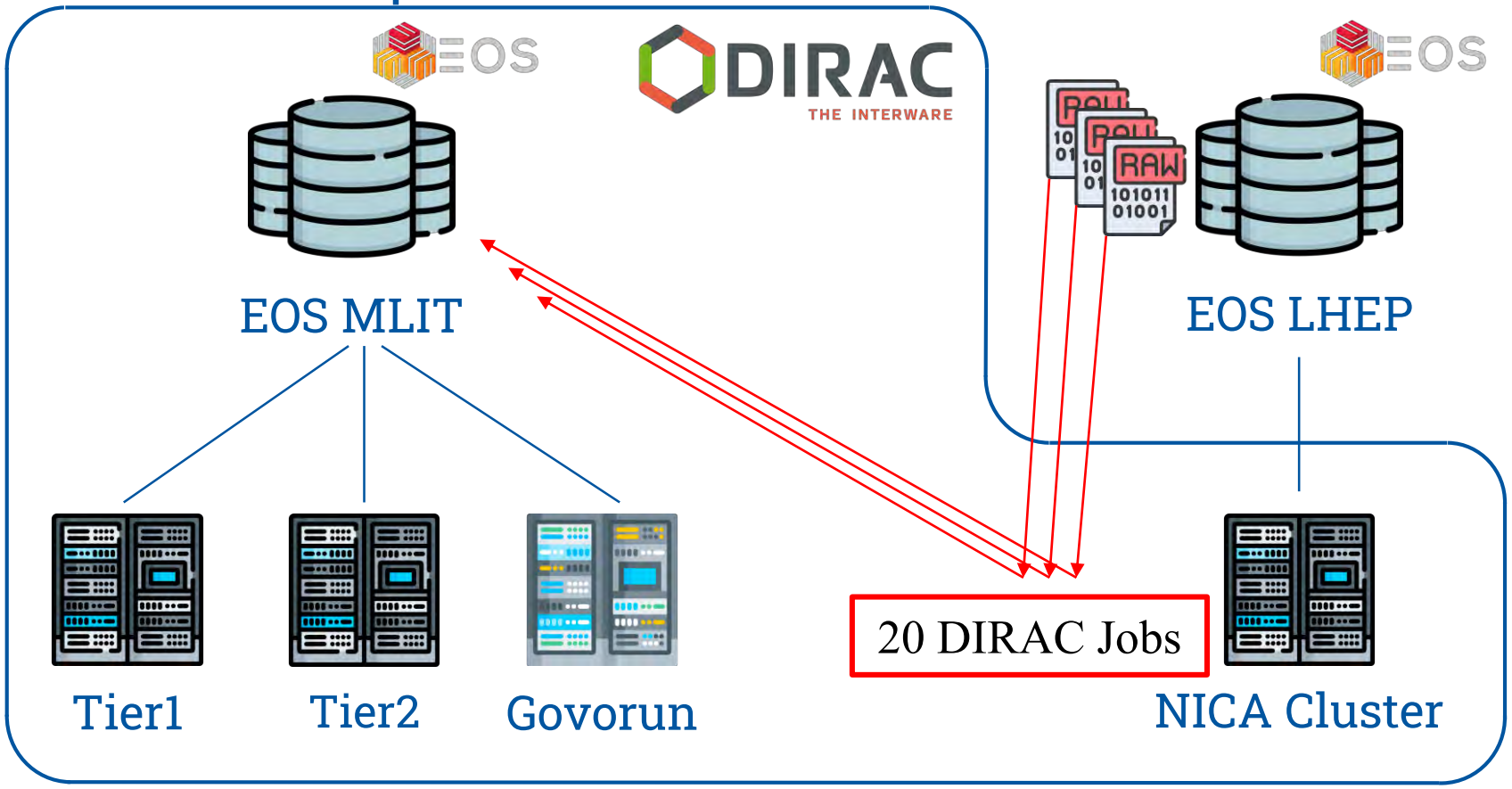


NRCN (National Research Computer Network) is the Russia's largest research and education (R&E) network. May allow execution of jobs submitted to Govorun on a resources of the network.
Massive tests with MPD jobs were performed successfully in the beginning of 2022

General scheme of resources



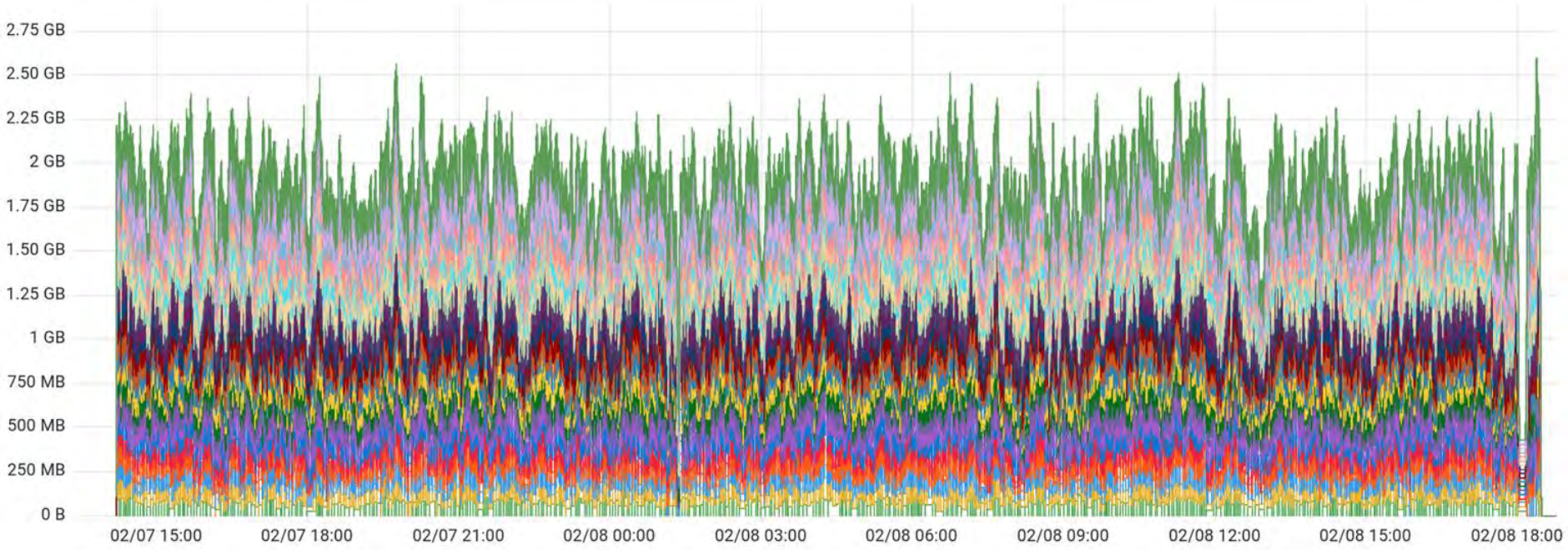
Step 1: Data transfer

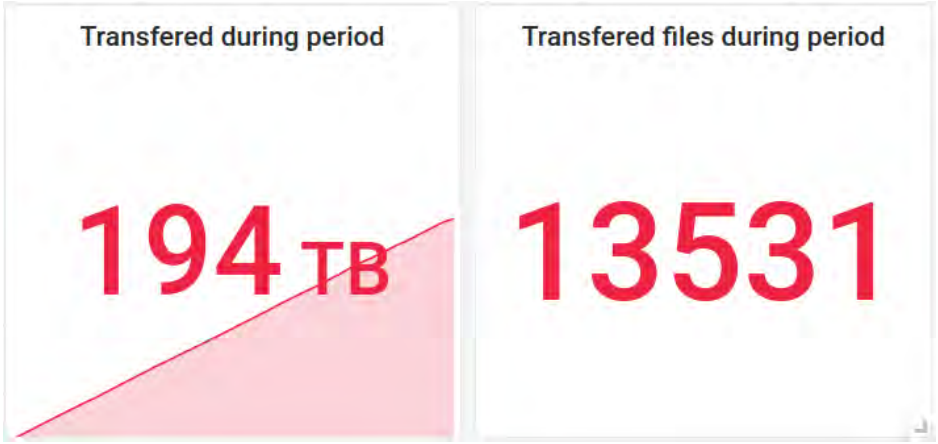


- Single stream of xrootd transfer can not exceed 100MB/s.
- NCX10* can sustain not more than 10 streams(1GB/s total). And that would overload its network.

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

Step 1: Data transfer





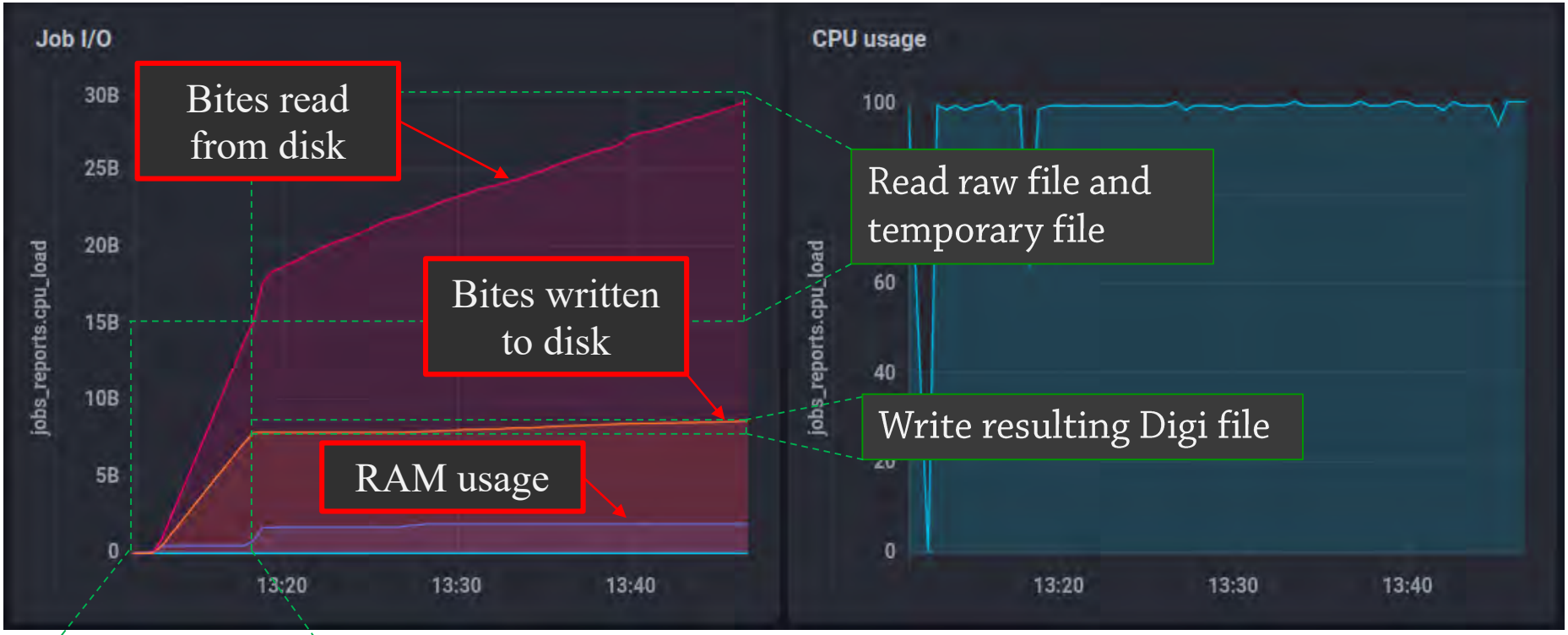
Average transfer speed on 20 streams
1.92 GB/s

Total transfer duration:
2d 15h

Step 2: Estimate the load

Size of files created during Run 8

Step 2: Raw2Digi job profiling

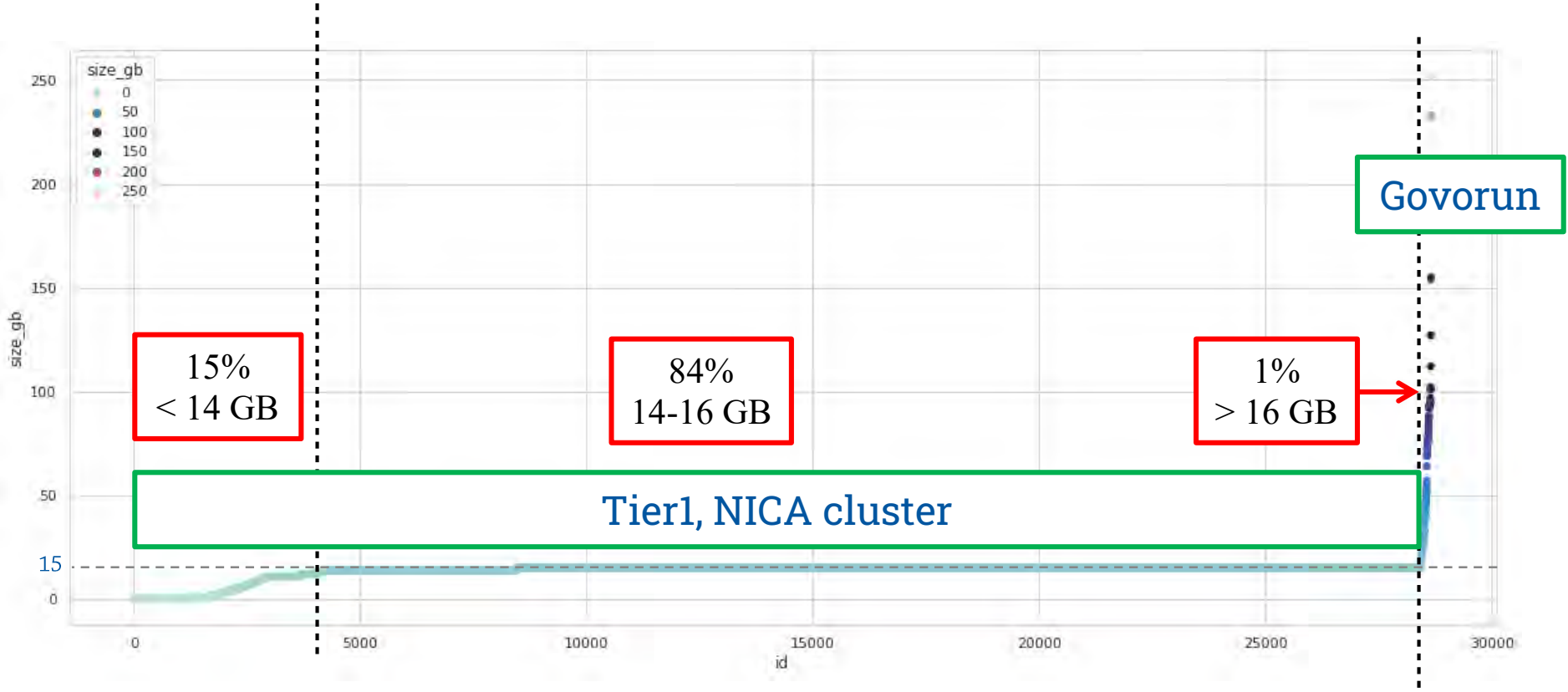


Initial read of 15GB raw file and creation of temporary 8 GB file

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

RAM usage : **~2GB**

Step 2: Raw file size sorted

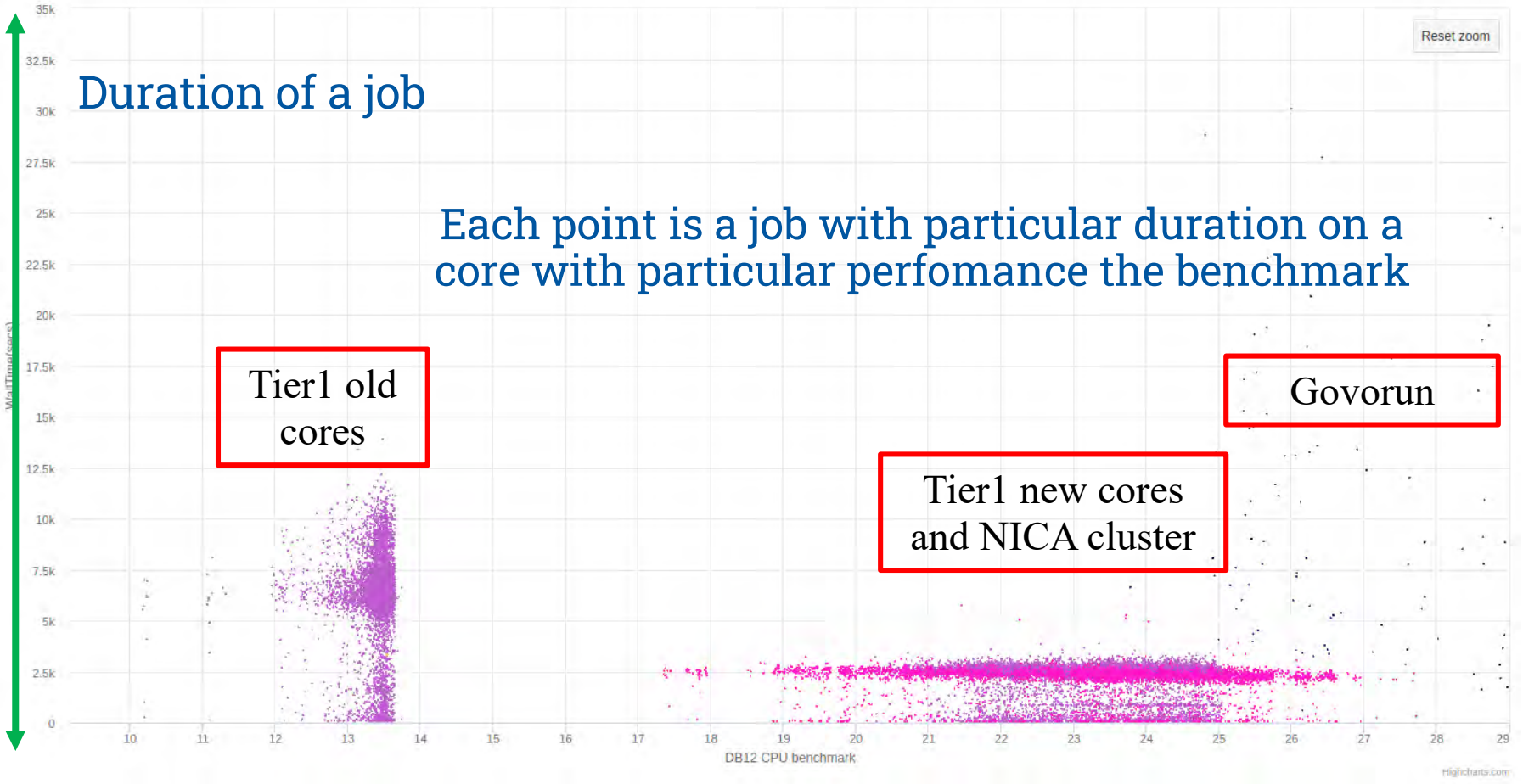


Only Govorun can efficiently process jobs with more than 30 GB of disk usage per one CPU core due to avalability of Luster storage attached to each worknode.

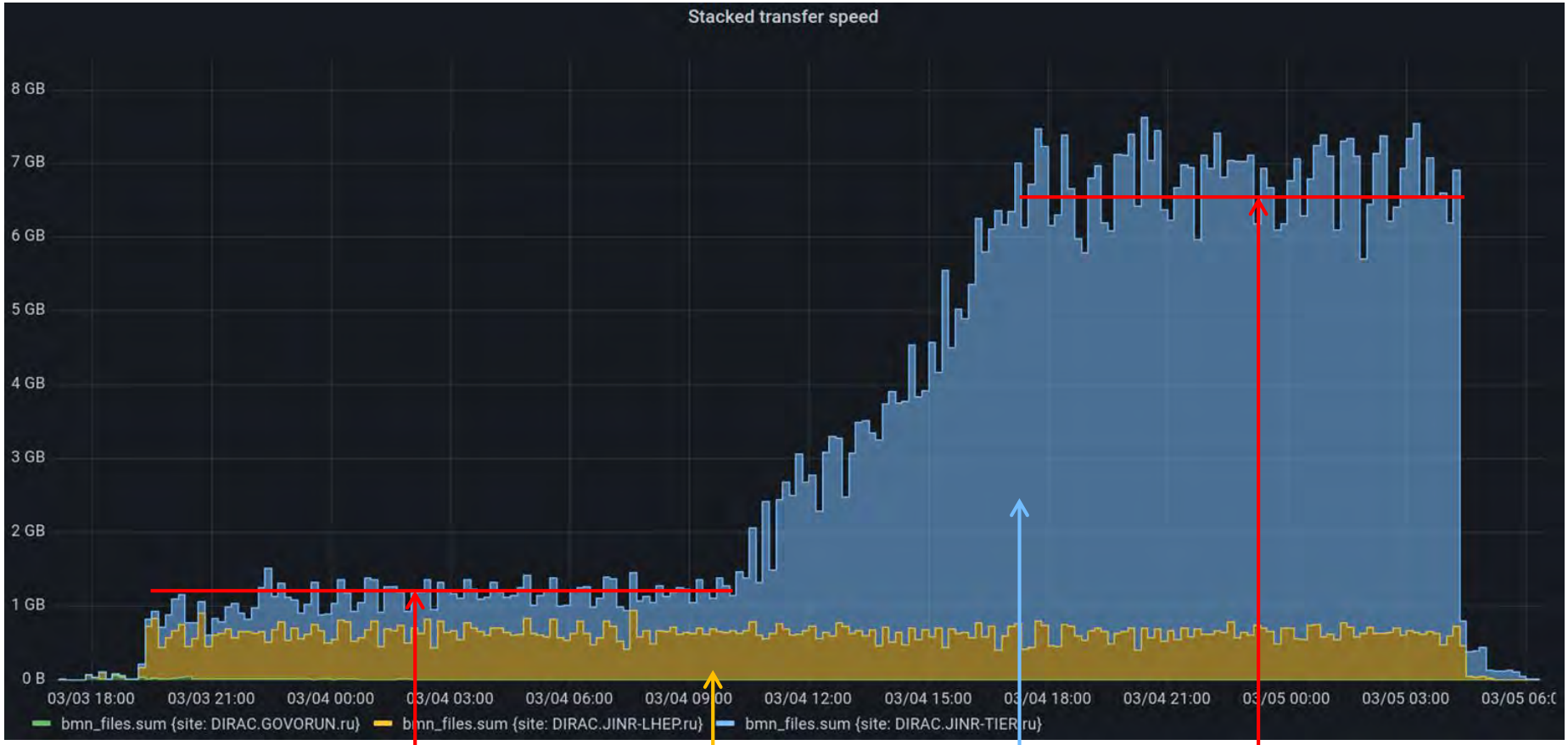
Luckly, in run 8 it is just 1% of all files that are larger than 16 GB. They can be processed on **Tier1** and **NICA cluster**(150 slots)

Step 3: Massive production Raw2Digi

Total duration of Raw2Digi campaign – 35 hours



Step 3: Network usage



300 jobs running
4 MB/s per job

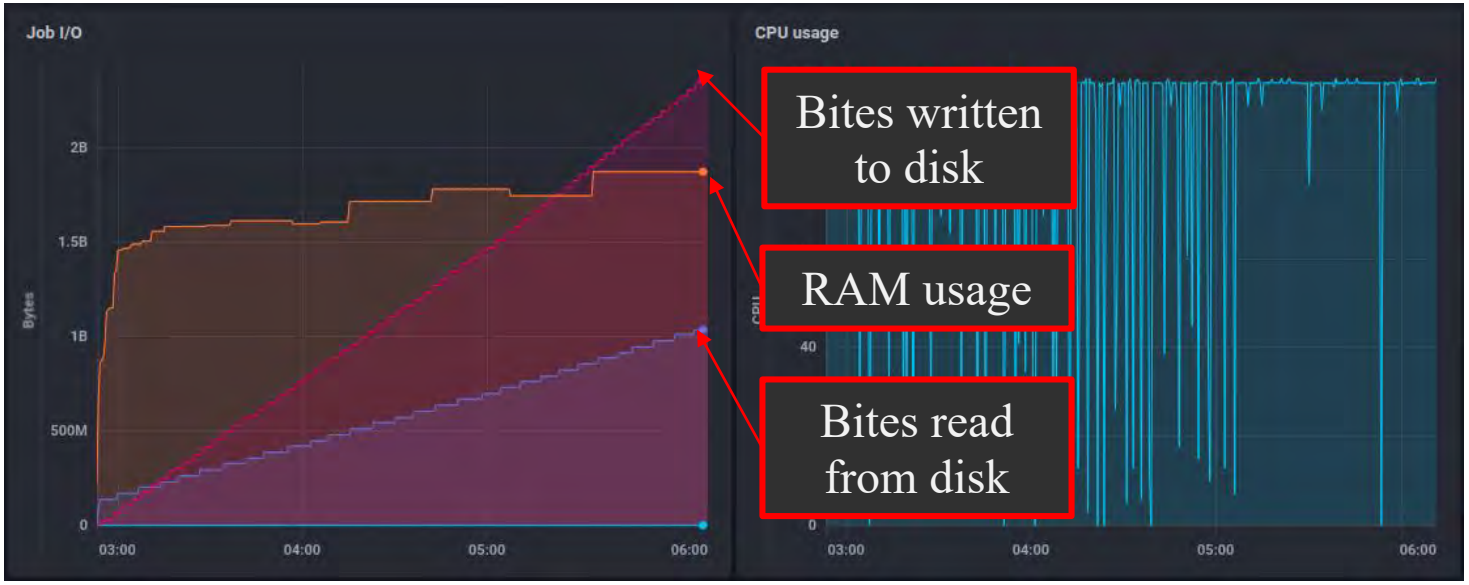
NICA Cluster

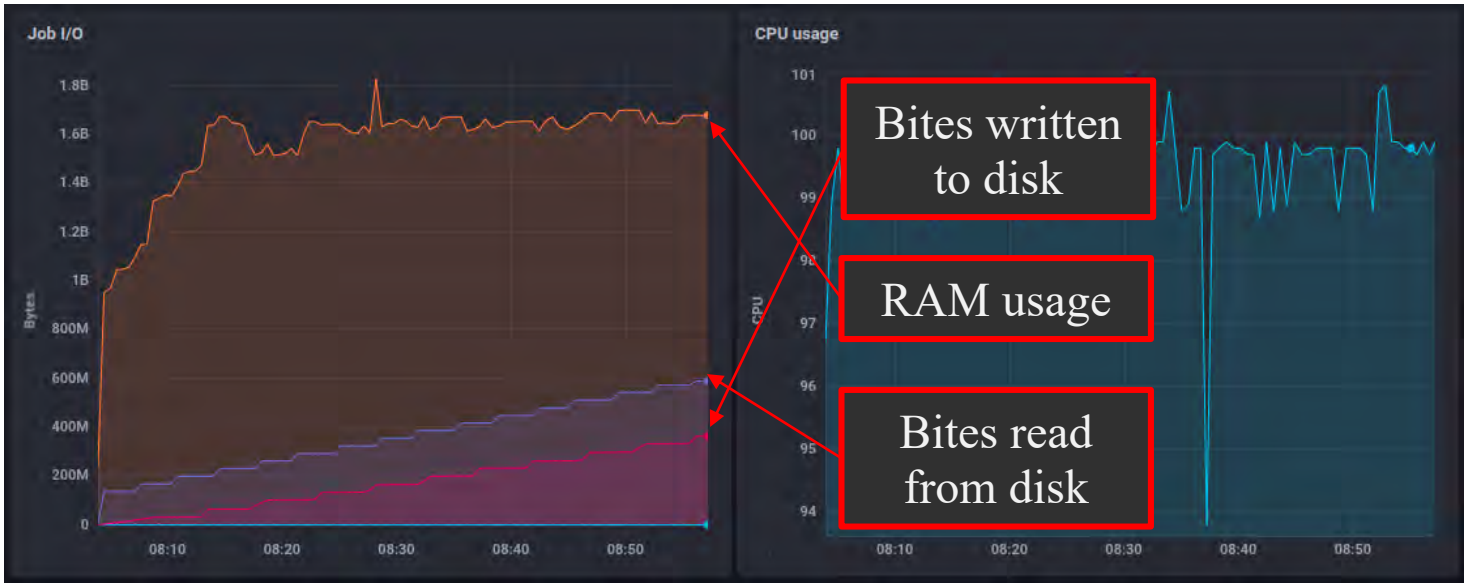
Tier1

1580 jobs running
4.1 MB/s per job

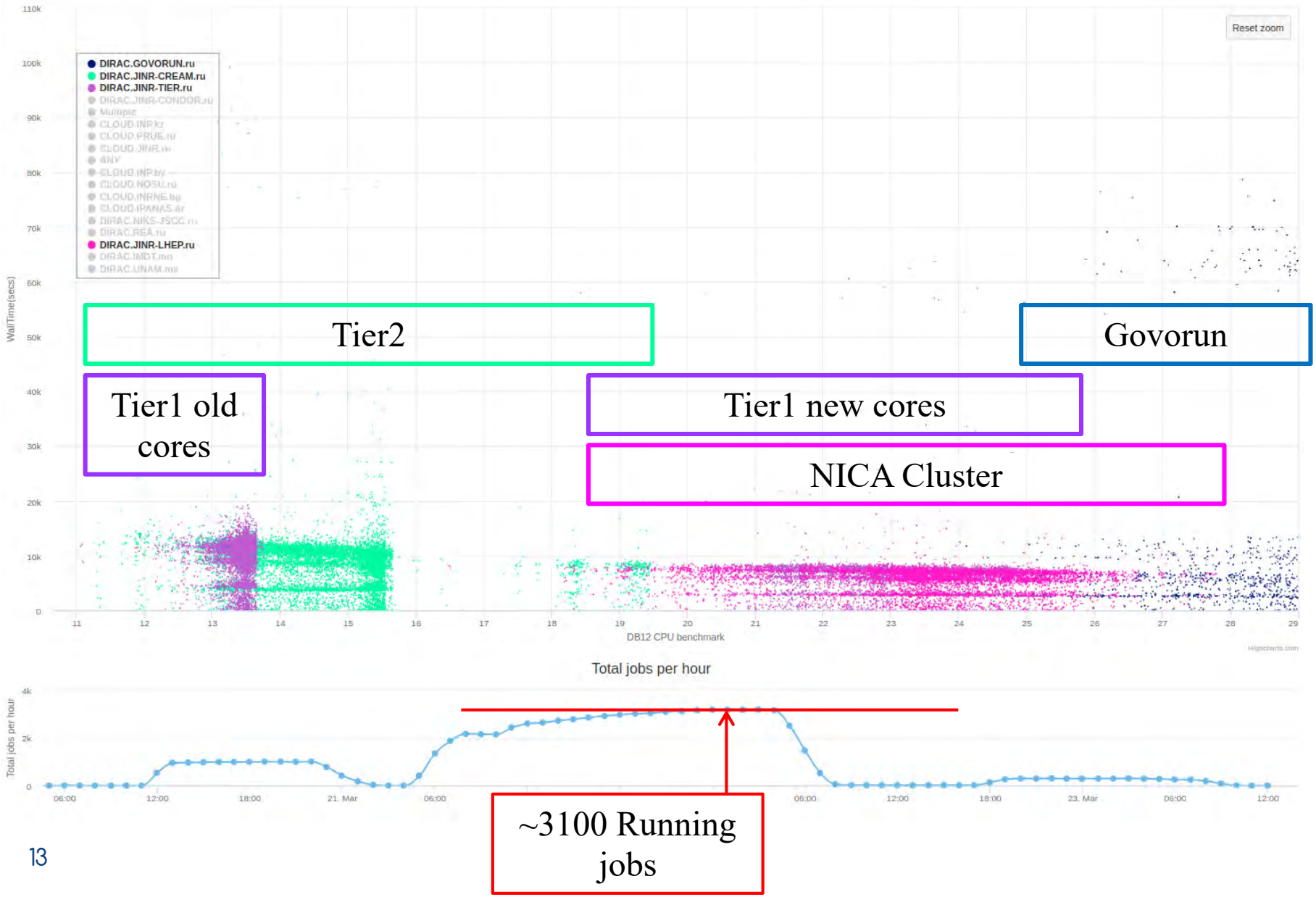
Maximal transfer speed (Read+Write) with EOS in MLIT – 7.5 GB/s

Step 4: Digi2Dst profiling

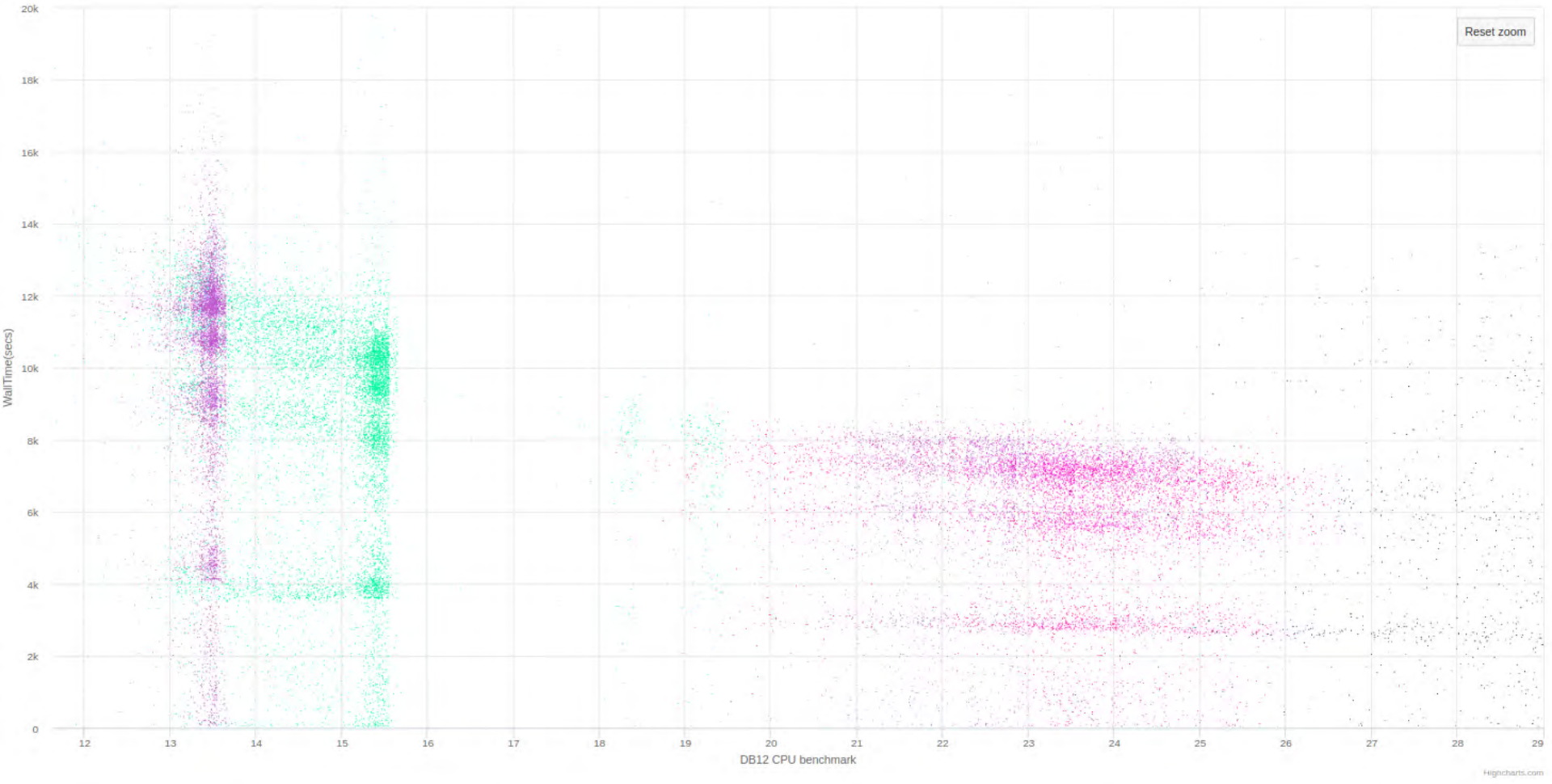




Step 5: Massive production Digi2Dst



Step 5: Digi2Dst job's segregation



Estimations

Max CPU cores utilized

Type	Tier1	Tier2	NICA cluster	Govorun	CPU core years required	Duration with all resources
Raw2Digi	1500	-	100	100	3.5 years	18 hours
Digi2Dst	1500	1000	300	200	10 years	30 hours

Disclaimer! All cores listed in this table may be used by any NICA experiment. That means that if another experiment use them at the same time the resources will be equaly divided between experiments.

Disclaimer! Assuming no major changes in bmnroot software. But they are gonna be there next time

Disclaimer! This numbers do not take into consideration “Cold start” problem. Real numbers may be larger by 3-6 hours

Results

- First time JINR computing infrastructure united by DIRAC was used for raw data reconstruction not in test mode but in production.
- Full BM@N run8 reconstruction takes considerable amount of resources. And, what's more important, specific computing resources that can sustain high load on disks. Up to now ~ 20 CPU core years has been consumed.
- Now, we have new experience. With all NICA computing resources available through DIRAC, presuming they are free from other's experiments work, it would take around 1 week to repeat all reconstruction.

List of participants

DIRAC: Igor Pelevanyk
BM@N: Konstantin Gertsenberger
Responsible for resources:
Govorun: Dmitry Podgainy, Dmitry Belyakov, Aleksandr Kokorev, Maxim Zuev,
NICA cluster: Ivan Slepov, Boris Schinov
Tier-1,Tier-2, EOS: Valery Mitsyn


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

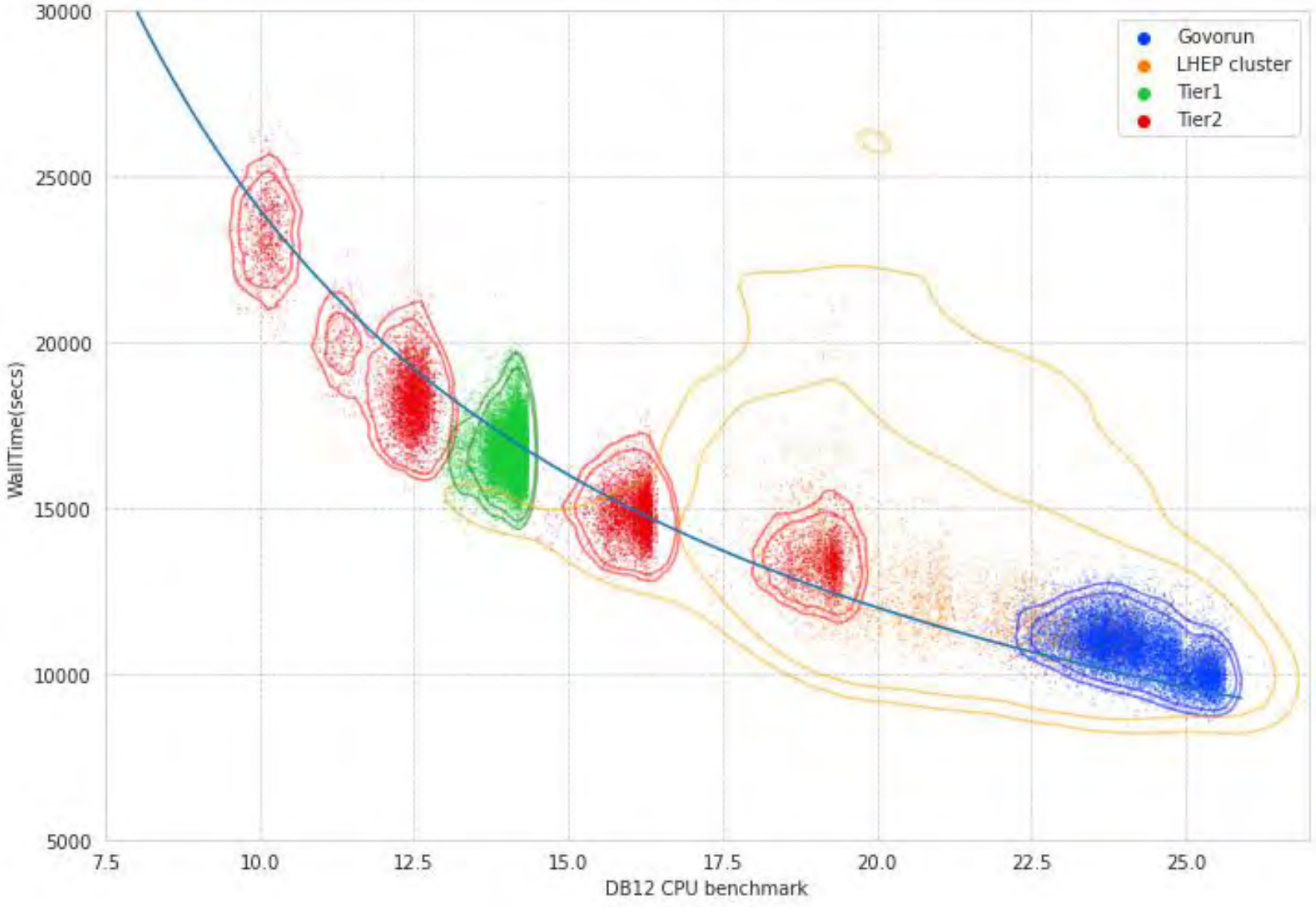
Piece of road from point A to B	←	A Monte-Carlo task
Speed of the car	←	Performance of the computer
Time to complete	←	Time to complete

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

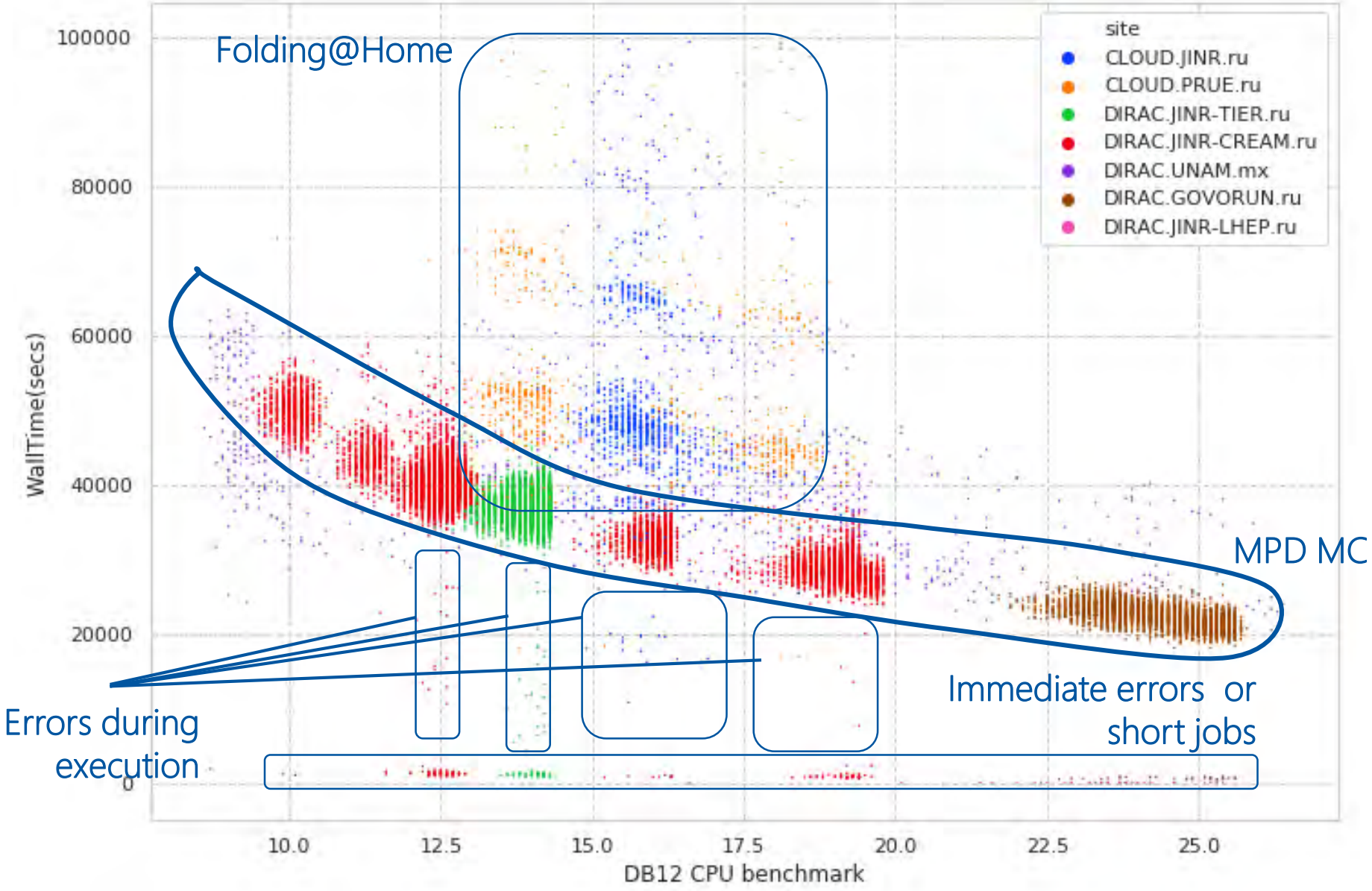
DB12 gives results like: 10(old slow core), 17 (standard server core), 27 (high performance core)

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.

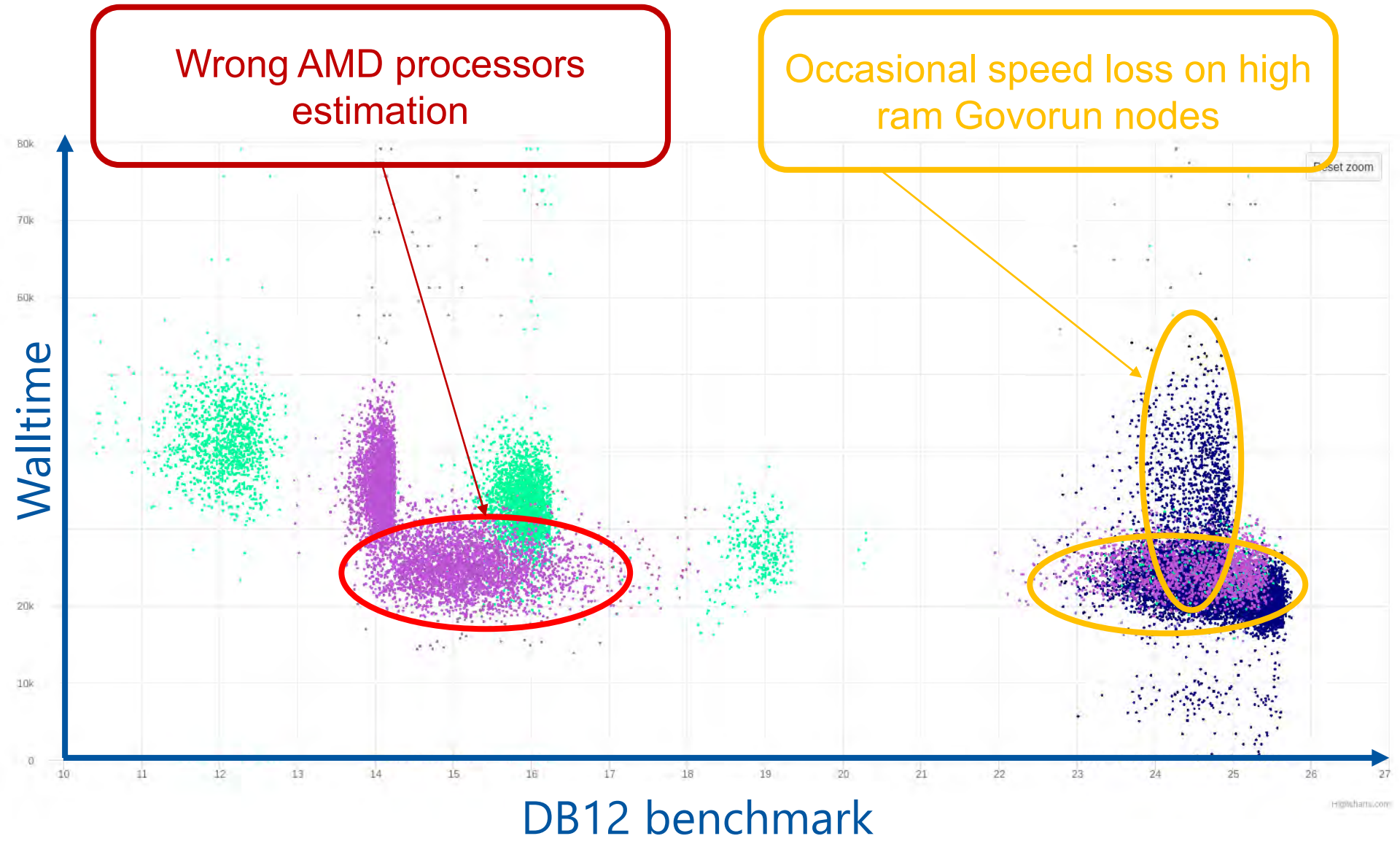
Performance analysis



Performance analysis



Discoveries



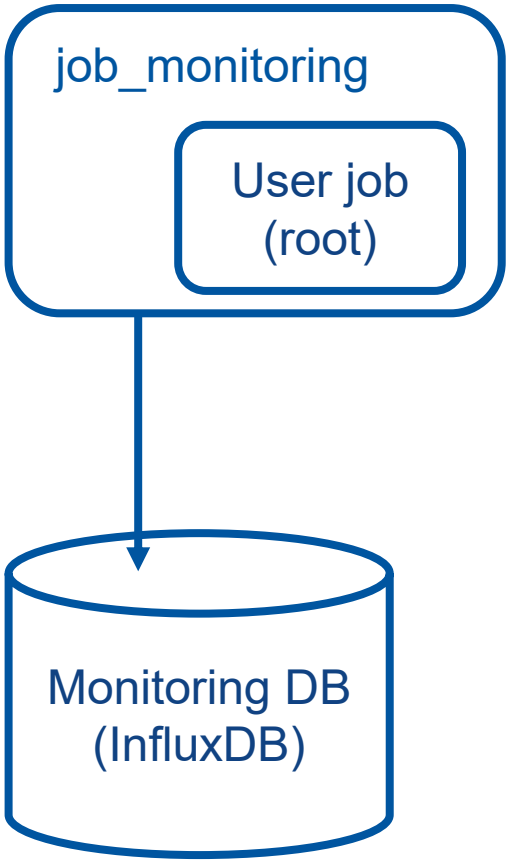
CPU core performance

Total CPU performance

User job monitoring

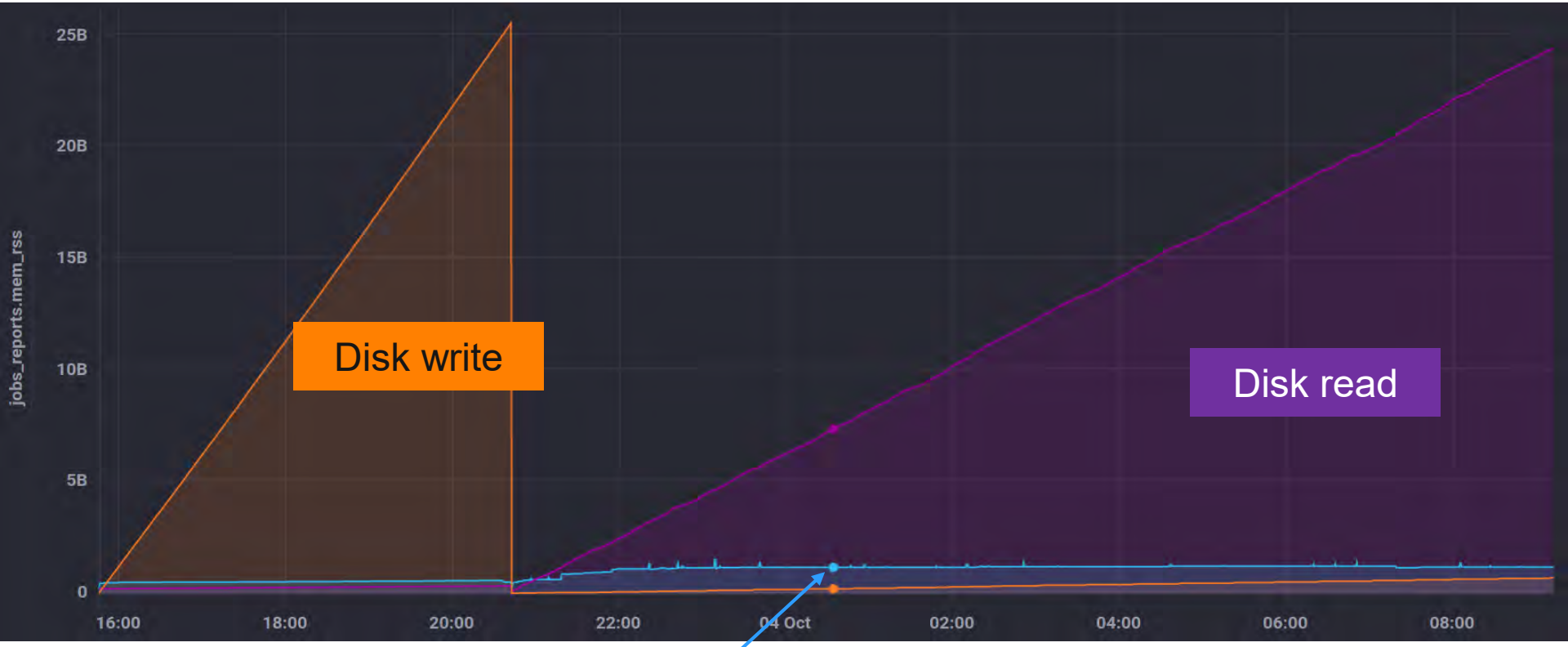
```
$ root macro.c(input)
```


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



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)