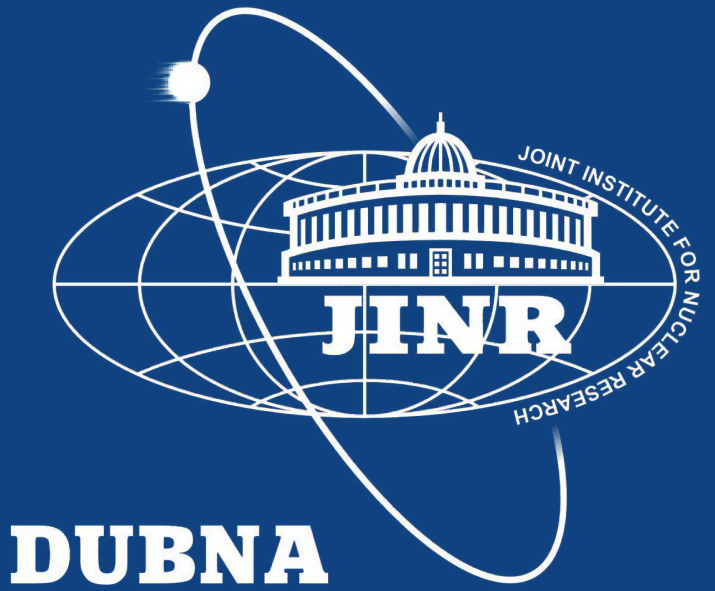

JOINT INSTITUTE FOR NUCLEAR RESEARCH
JINR
DUBNA

BM@N mass data production on the distributed computing infrastructure with DIRAC

Konstantin Gertsenberger, Igor Pelevanyuk

15 September 2022

A moment of retrospective

6th BM@N collaboration meeting:

General information about DIRAC and how it may be useful for BM@N tasksz

7th BM@N collaboration meeting:

Initial tests have been successfully performed. Basic workflow and scripts designed and presented.

8th BM@N collaboration meeting:

First real(useful) Monte-Carlo have been performed!

9th BM@N collaboration meeting:

More Monte-Carlo have been done. Preparing for raw data processing

File types

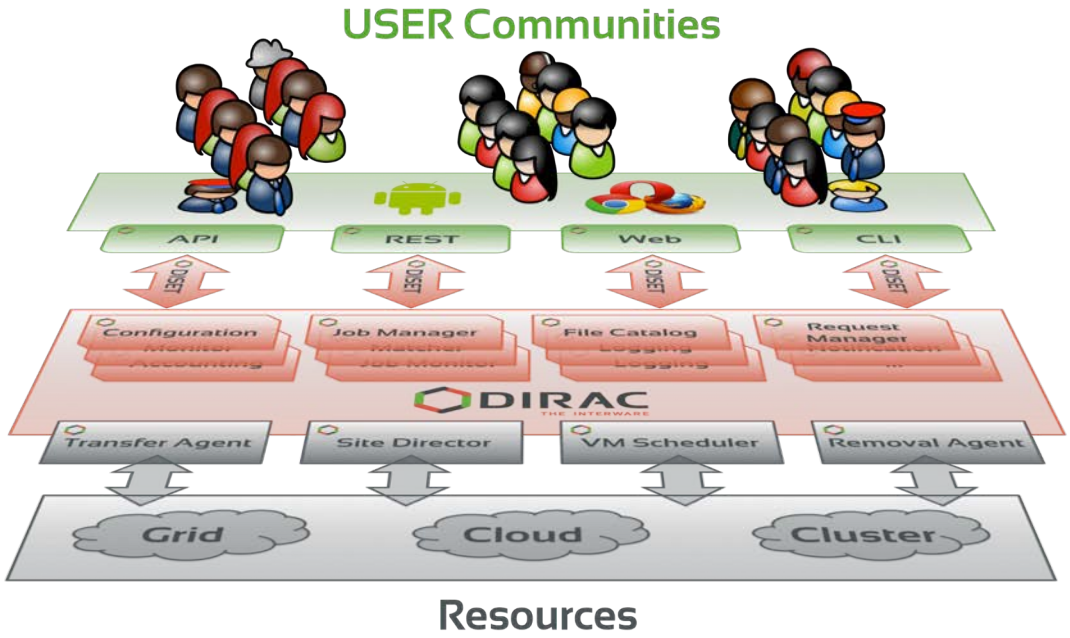
Type	Decription
RAW	Raw data of events written by DAQ after Event Builder
DIGI(digits)	Digits of detectors after digitizer macros
DST_{exp}	Reconstructed events: hits, tracks, vertex and other reconstructed data
GEN	Generated events after simulation
DST_{sim}	Reconstructed events containing modeled information for comparison

BMN offline job types

Experiment data	Monte-Carlo data
RawToDigit	GenToSim
DigitToDst	SimToDst
DstToAna	DstToAna

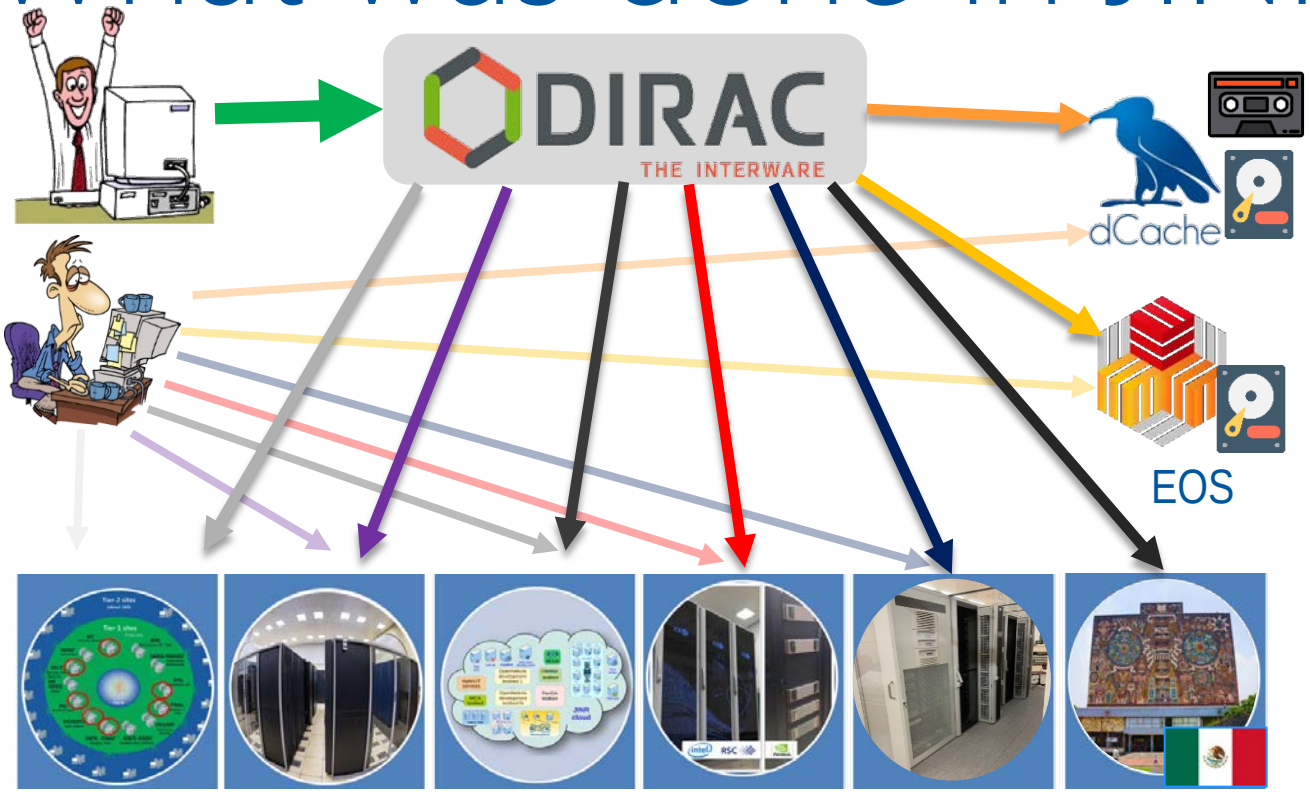
What is DIRAC?

DIRAC provides all the necessary components to build ad-hoc grid infrastructures **interconnecting** computing resources of different types, allowing **interoperability** and simplifying **interfaces**. This allows to speak about the DIRAC *interware*.



- Web
- CLI
- API
- REST

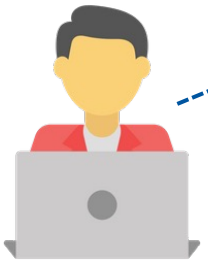
What was done in JINR



Tier-1	CICC/Tier-2	Clouds	Govorun	NICA Cluster	UNAM
Running	**Running**	**Running**	**Running**	**Running**	**Running**

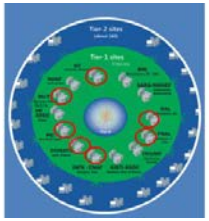
The computing resources of the JINR Multifunctional Information and Computing Complex, clouds in JINR Member-States, cluster from Mexico University were combined using the DIRAC Interware.

Workload management



Submit thousand of jobs to DIRAC Job Queue

DIRAC
THE INTERWARE


Tier-1

CICC/Tier-2

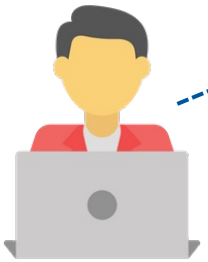
Clouds

Govorun

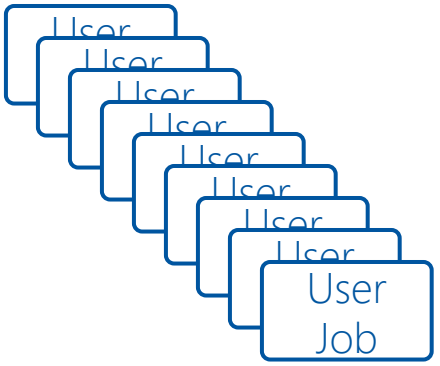
NICA Cluster

UNAM

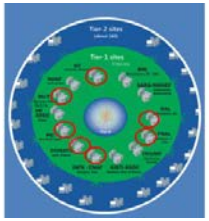
Workload management



Submit thousand of jobs to DIRAC Job Queue



DIRAC
THE INTERWARE


Tier-1

CICC/Tier-2

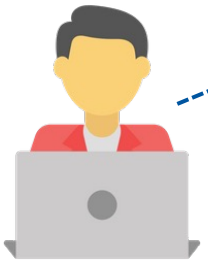
Clouds

Govorun

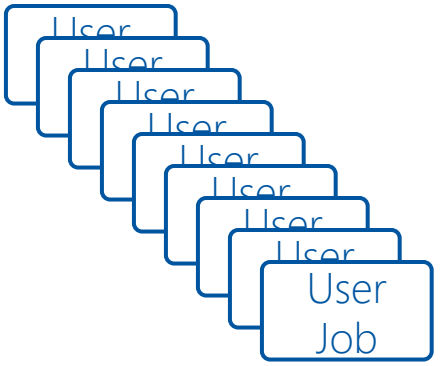
NICA Cluster

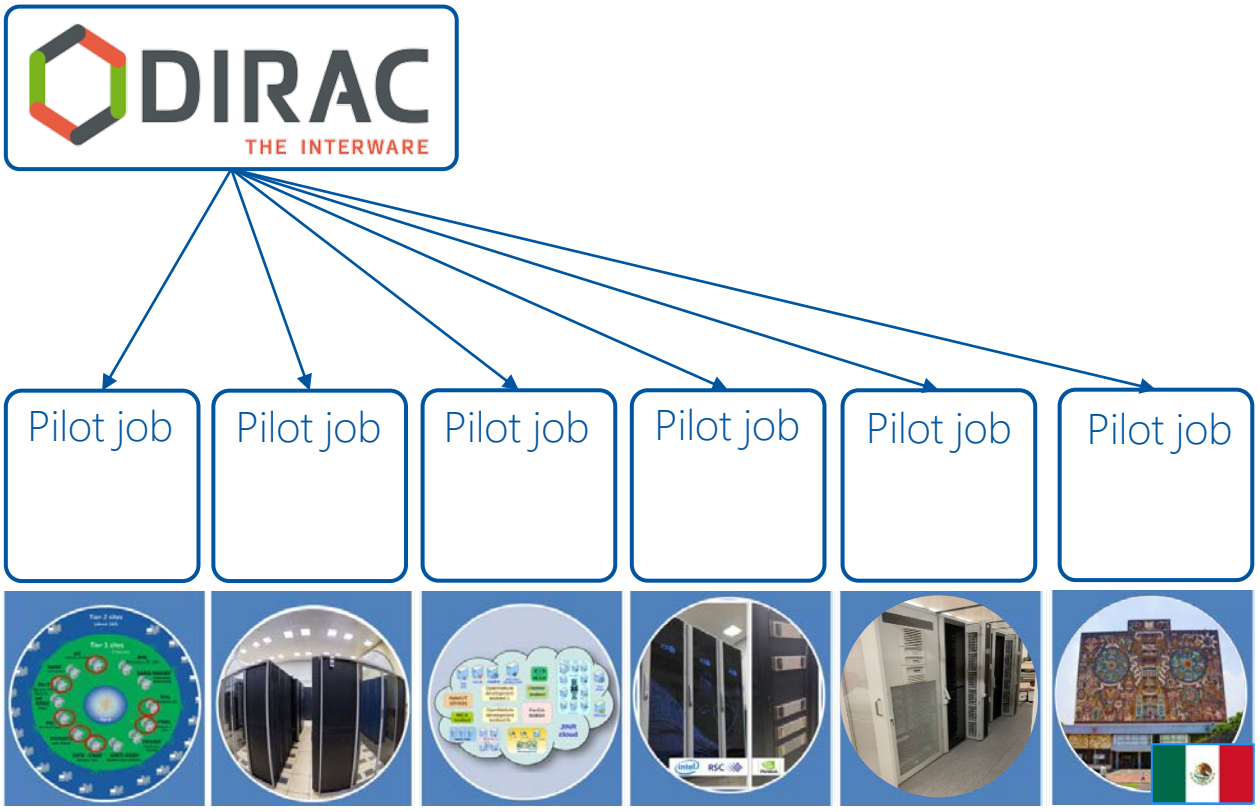
UNAM

Workload management

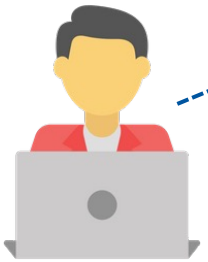


Submit thousand of jobs to DIRAC Job Queue





Workload management

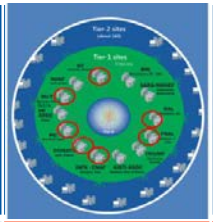


Submit thousand of jobs to DIRAC Job Queue

DIRAC
THE INTERWARE

Tier-1 | CICC/Tier-2 | Clouds | Govorun | NICA Cluster | UNAM

DIRAC: Jobs vs Resources

	Tier-1	CICC/Tier-2	Clouds	Govorun NICA Cluster	
					
RawToDigit			Only with CVMFS		
DigitToDst			Only with CVMFS		
GenToSim			Only with CVMFS		
SimToDst			Only with CVMFS		

This schema only relates to DIRAC jobs, not jobs submitted directly to resources

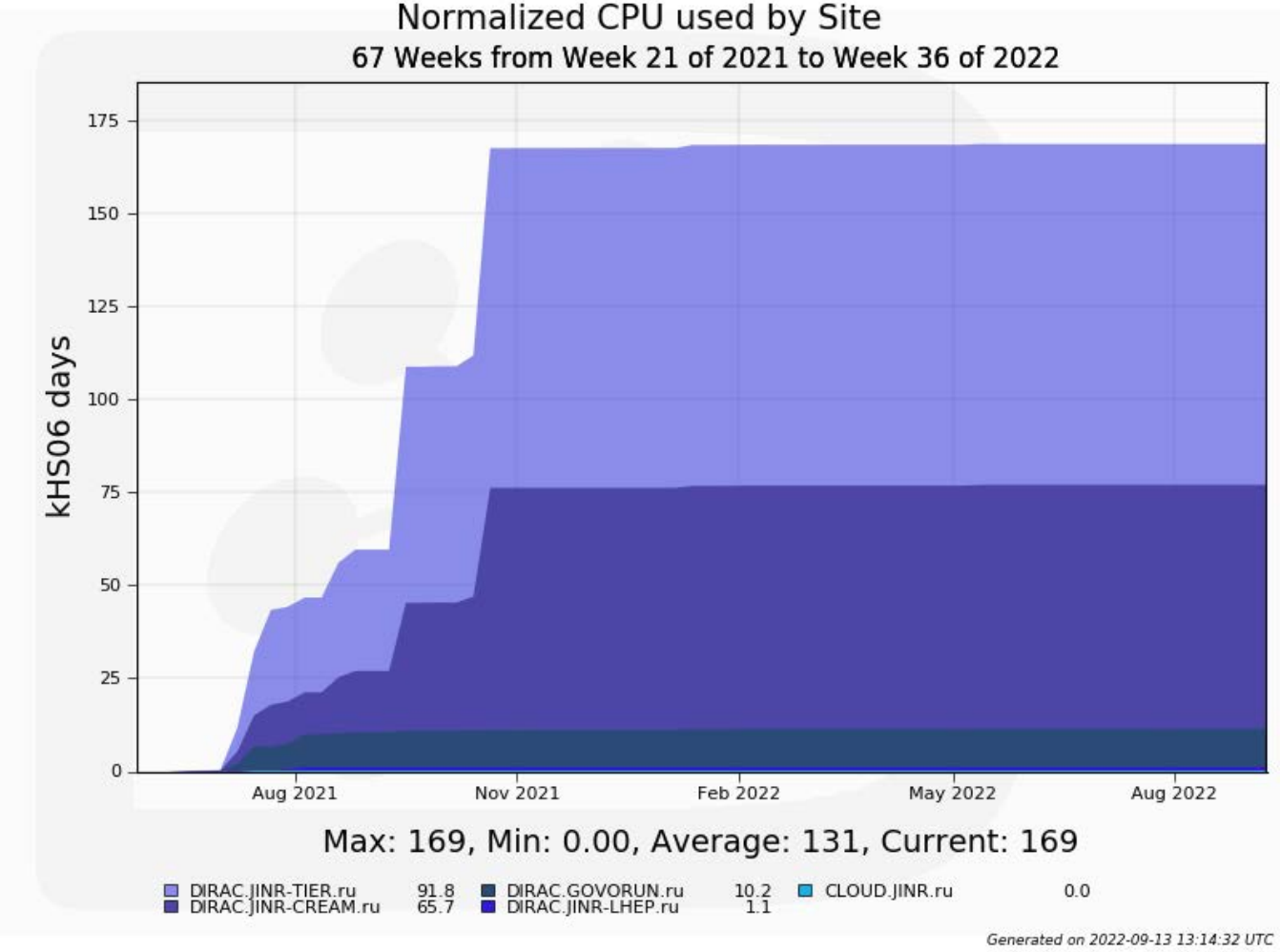
Explanation

RawToDigi – Huge size of input files is main limitation. There should be at least 100 GB of disk space per CPU core.

Cloud resources – potentially can perform all workload. Placement of software in CVMFS is essential. But the network may be a limiting factor for remote clouds.

NICA cluster – was used less in the last half a year by all DIRAC users(BM@N, MPD, SPD).

BM@N statistic



Tier1

Tier2

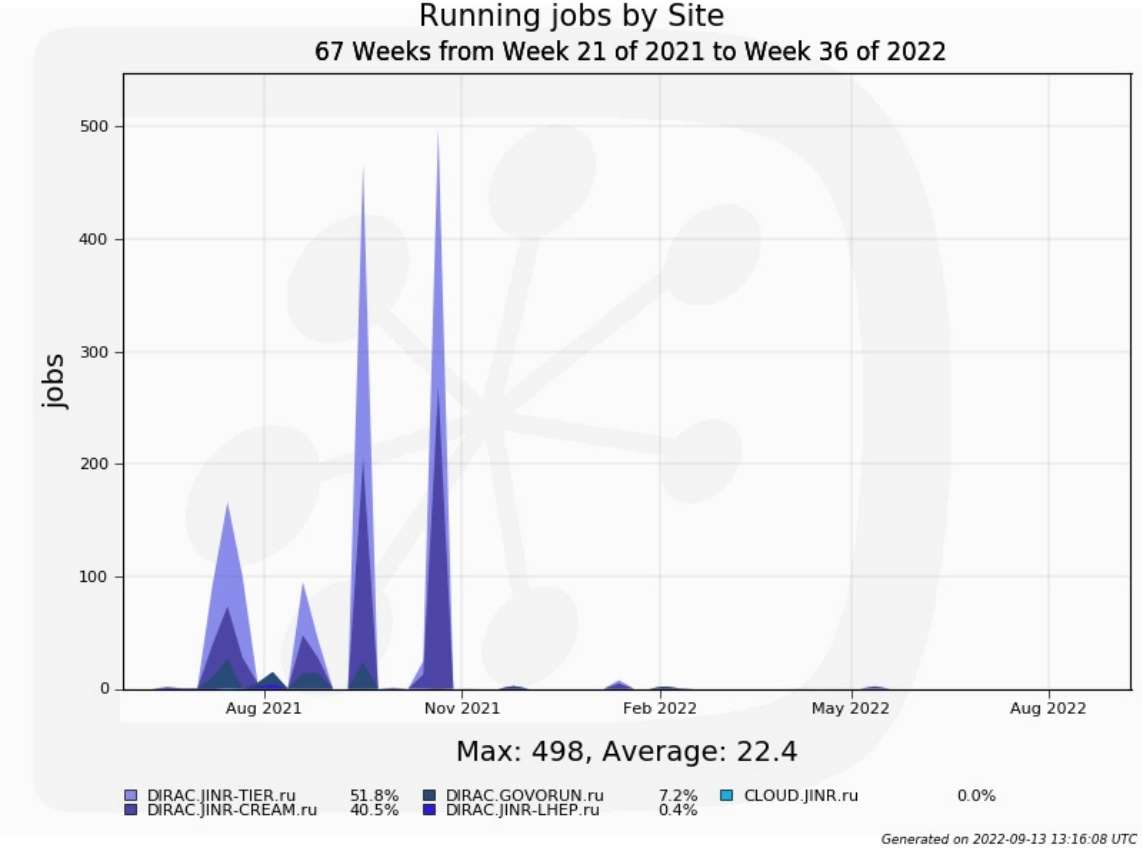
Govorun

Total number of jobs: 18,900

Total wall time: 29 years

Average duration: 13 hours

Available resources



In the mid of September total amount of running jobs exceeded 1600.

Quotas(cores):
Tier1: 920 (NICA shared)
Tier2: 1000 (NICA shared)
Govorun: 192
NICA cluster: 250
JINR Cloud: 90 (All shared)
Members-states clouds: ~500 (All shared)

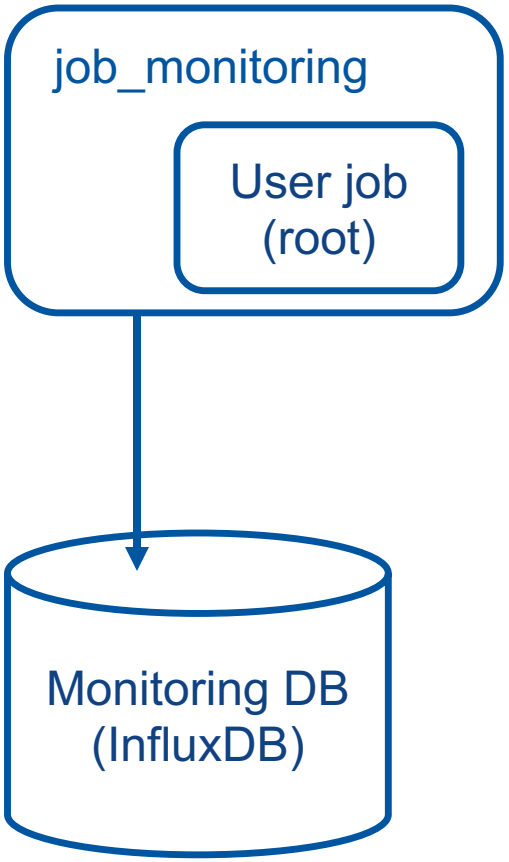
BMN Raw->Digi workflow

1. Check the resource and configure software
2. Download RAW file
3. Perform BmnDataToRoot.C
4. Copy result_digi.root to MLIT-EOS over DIRAC

User job monitoring

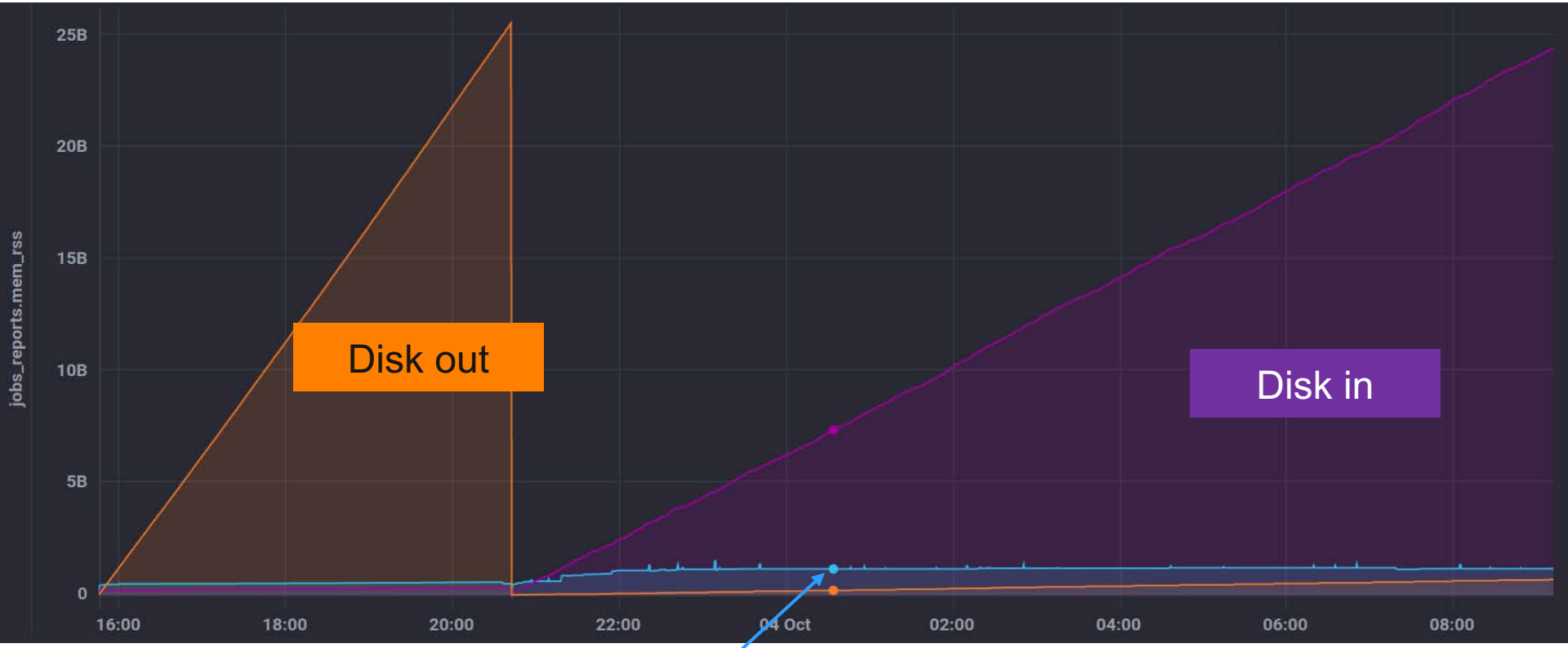
```
$ root macro.c(input)
```

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



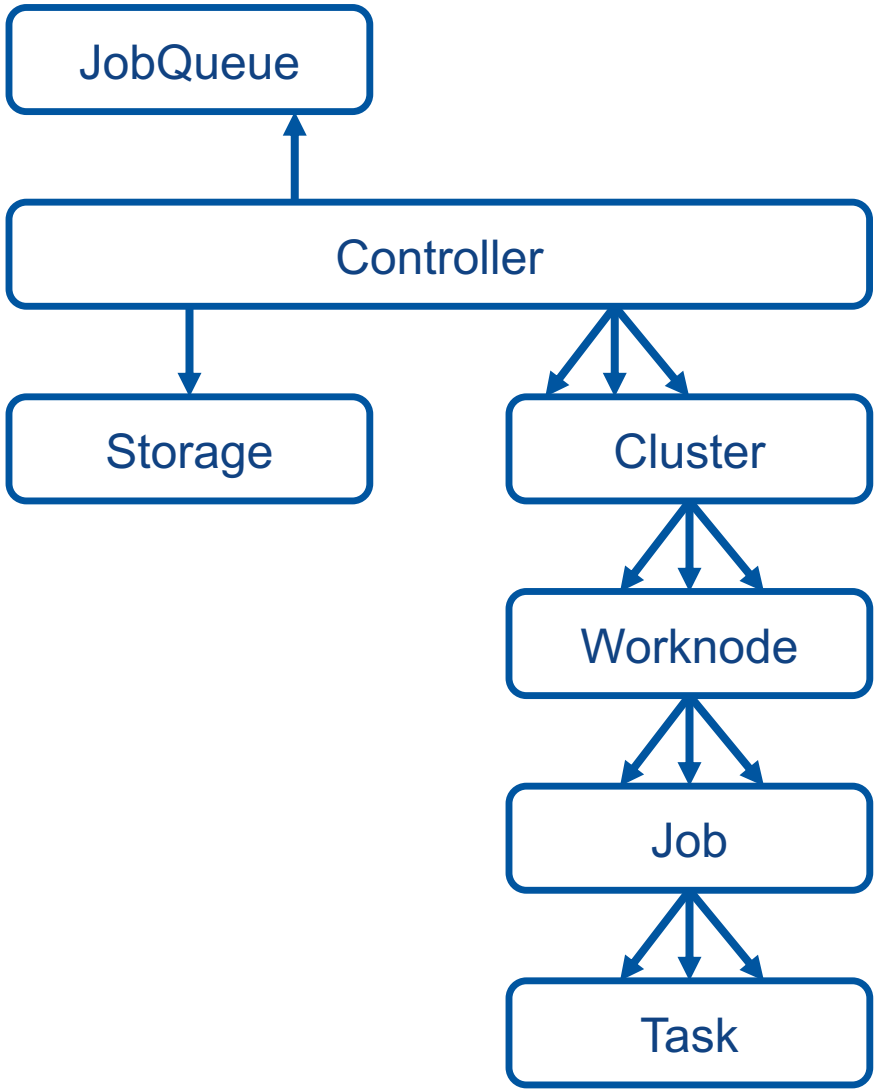
User job monitoring

GenToDst job on Govorun



DIRAC load prediction

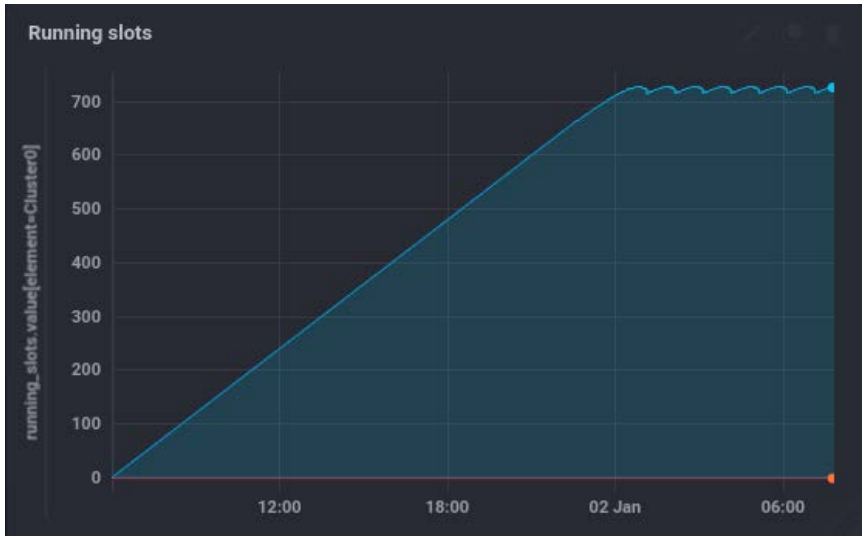
- Written in python to predict CPU, RAM, network and disk load
- Uses data about performance of resources integrated in DIRAC
- It is used to check the behavior of DIRAC jobs in real infrastructure.
- InfluxDB is used for results storage and visualization



BM@N computing parameters

- If we have 20 Govorun worknodes, 40 cores available on each worknode.
- If 100 MB/s maximum disk writing speed on each worknode.
- If new 40 GB RAW file appears every 90 seconds. 105000 events in each RAW file.
- If each event processing time is 0.5 sec – one file processing will last for 14.5 hours.

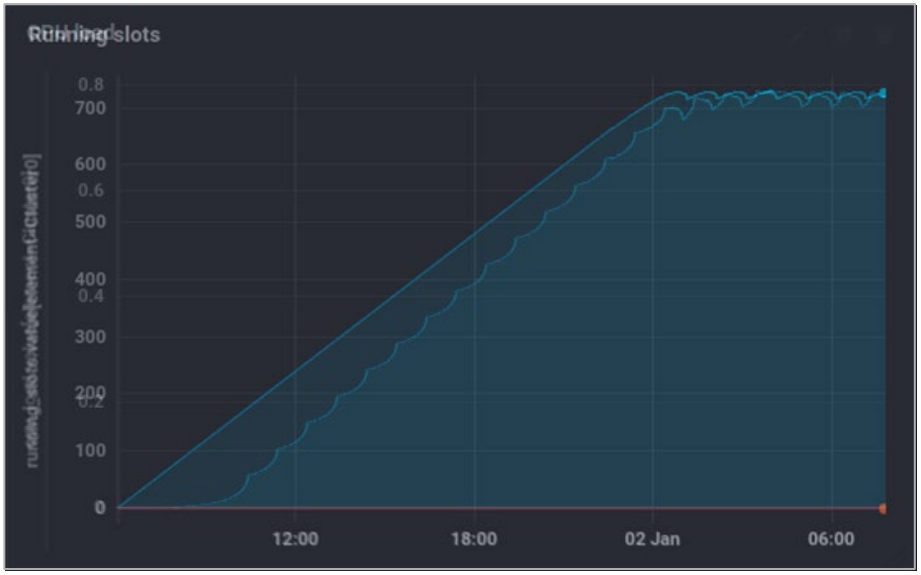
Results



- No disaster happening
- Stable state of the system after ~19 hours
- Maxmim amount of running jobs ~730 (91% of slots)
- Average CPU load of available Govorun resources ~ 80%.

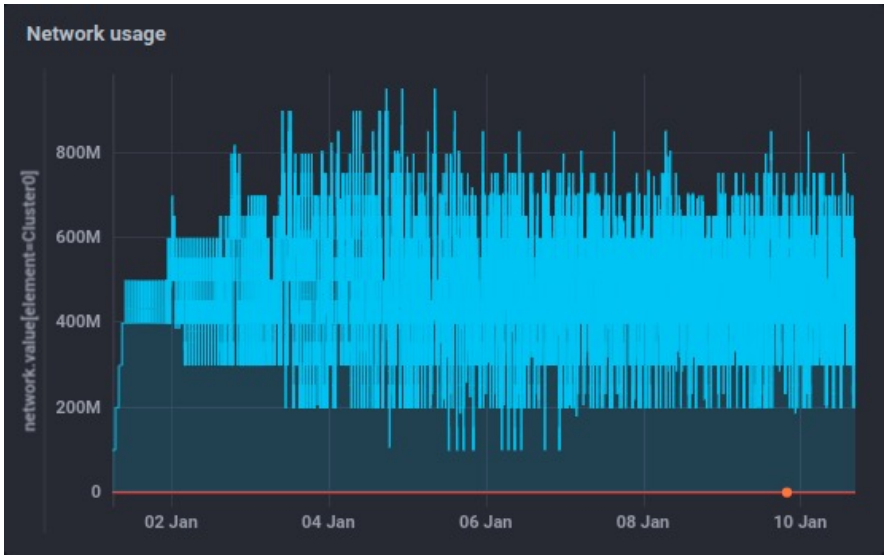
- Network usage not more than 700 MB/s
- Average usage between 400 and 500 MB/s

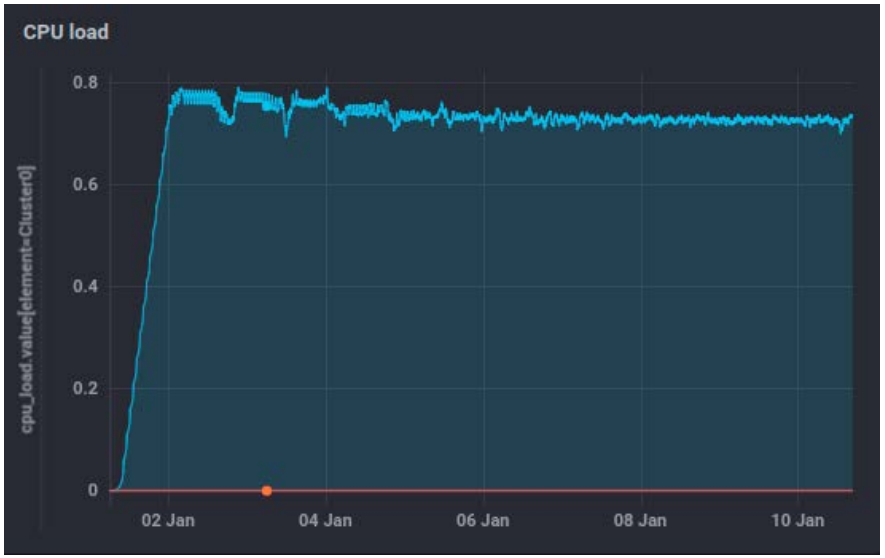
Strange results

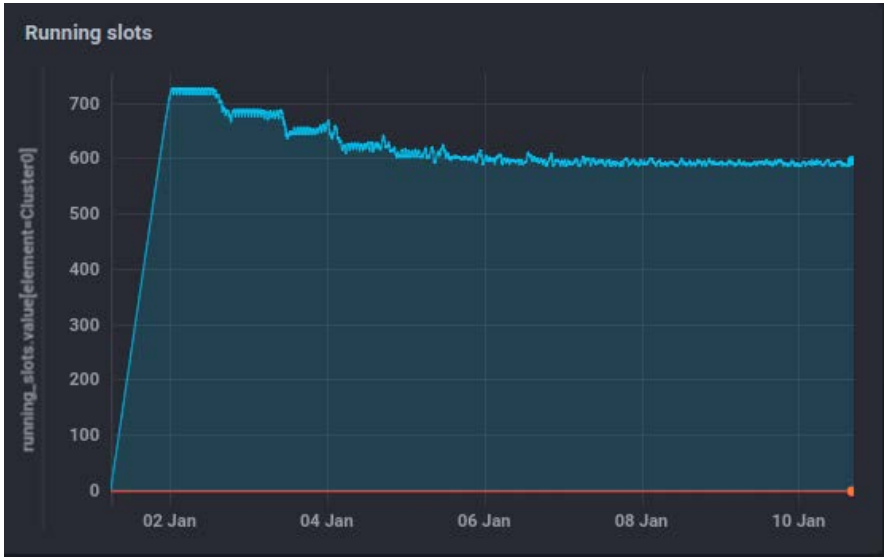


- These “waves” are on CPU load graph is definitely an issue.
- If we place Running slots graph on top of CPU load graph we see that in the beginning many slots occupied by jobs which struggle to download data.
- What if we will distribute jobs among worknodes randomly?

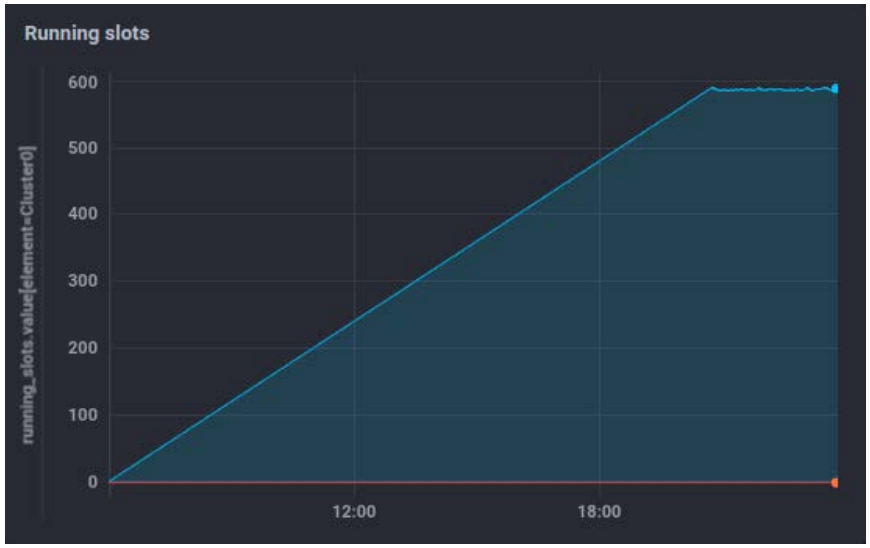
Results


Network usage
network.value[element=Cluster0]
800M
600M
400M
200M
0
02 Jan
04 Jan
06 Jan
08 Jan
10 Jan


CPU load
cpu_load.value[element=Cluster0]
0.8
0.6
0.4
0.2
0
02 Jan
04 Jan
06 Jan
08 Jan
10 Jan


Running slots
running_slots.value[element=Cluster0]
700
600
500
400
300
200
100
0
02 Jan
04 Jan
06 Jan
08 Jan
10 Jan

Results

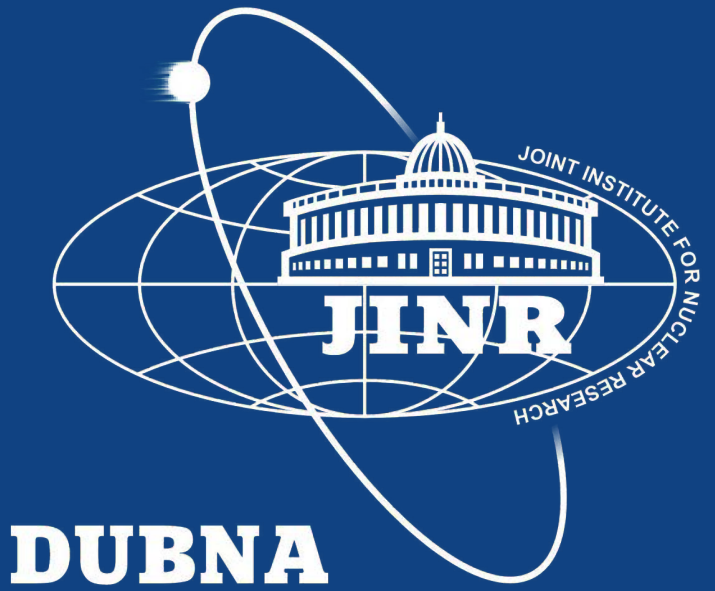


- No disaster happening
- Stable state of the system after ~15 hours
- Maxmim amount of running jobs ~591 (74% of slots)
- Average CPU load of available Govorun resources ~ 73%.

- Network usage not more than 700 MB/s
- Average usage ~450MB/s

Conclusion on BM@N+DIRAC

- If our estimations are correct. DIRAC + Govorun may be used for Raw->Digi jobs during the whole BM@N run.
- If we use DIRAC for Raw->Digi, Govorun is the only computing resource that can effectively provide more than 100GB of disk space per running CPU core.
- Within DIRAC Tier1, Tier2 are the best resources for MC and Digi->Dst jobs. Clouds are also available and may be used if software is in CVMFS.
- BM@N root is better to be executed from CVMFS.


JOINT INSTITUTE FOR NUCLEAR RESEARCH
JINR
DUBNA