The Analysis Facility at GSI
One year of operation under analysis load

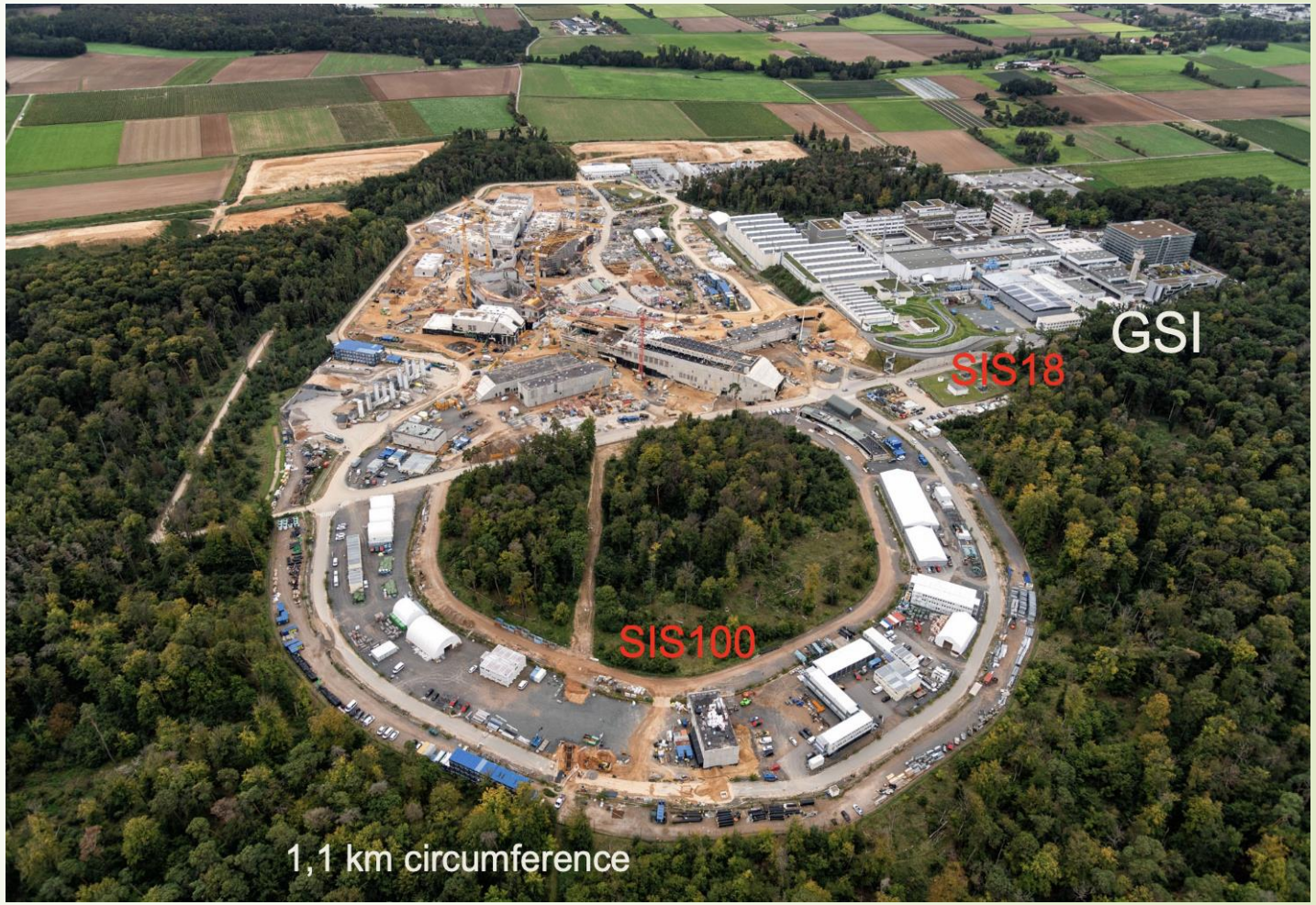
Sören Fleischer, Raffaele Grosso and Mohammad Al-Turany

Computing at GSI/FAIR: Where are we going ?

- Computing at GSI
  - Where we are and where we are going
- The GSI Analysis Facility
  - Site resources
  - GSI queues
  - Solved and open issues
  - AF requirements, current state
  - Plans

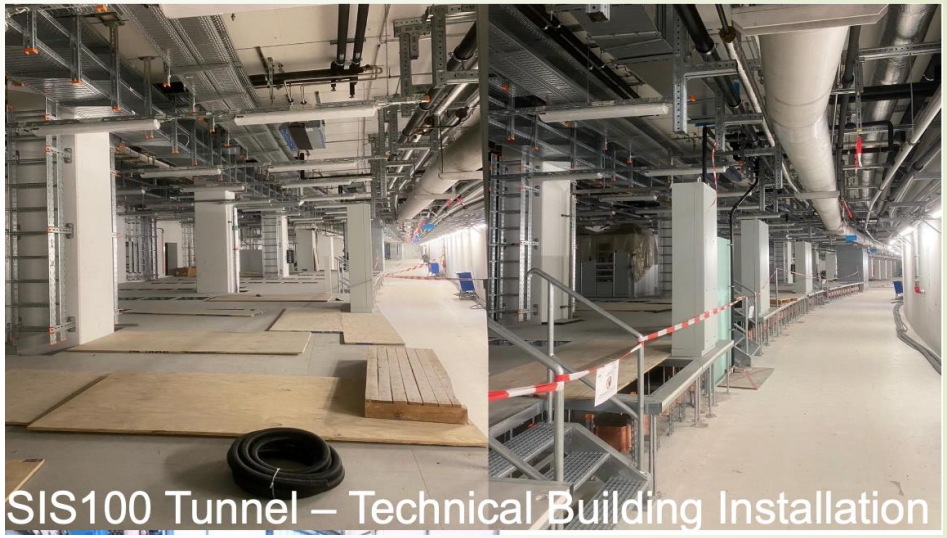
Facility for Antiproton and Ion Research in Europe

- 3000 scientists from 50 countries
- First experiments expected in 2028

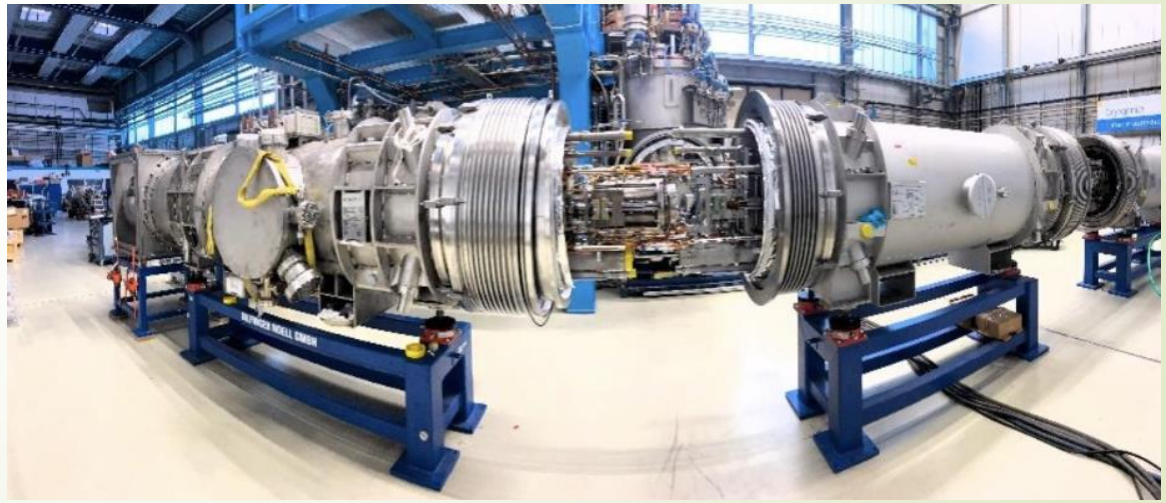


Highlights from FAIR Construction Site – installation started 2024

First power supply unit

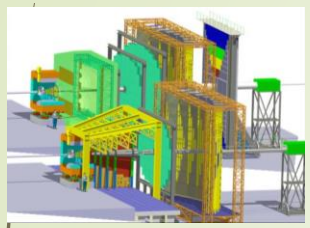


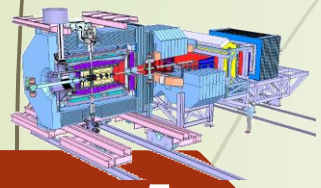
SIS100 Tunnel – Technical Building Installation



First thermal cycle of the SIS100 string

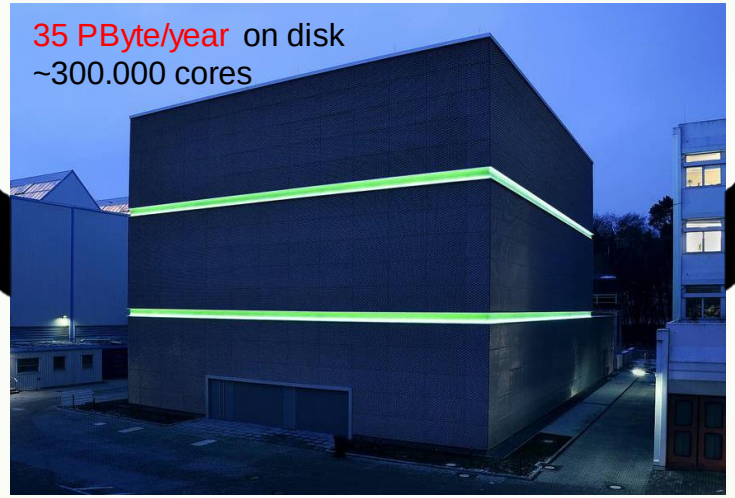
Dynamically allocated resources for exclusive usage and limited time





5

Computing at FAIR: The resources in the Green-Cube will be shared between the different FAIR/GSI Partners



Analysis Facilities (Grid Tier2)

ALICE

No separate hardware for the online clusters of the FAIR experiments

Generic batch farm for GSI/FAIR Users

CBM

R3B

panda

HADES

AGATA
ADVANCED GAMMA TRACKING ARRAY

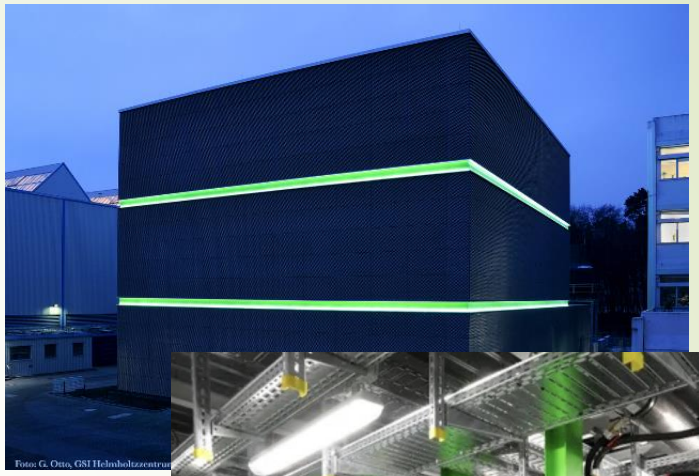
BIOMAT

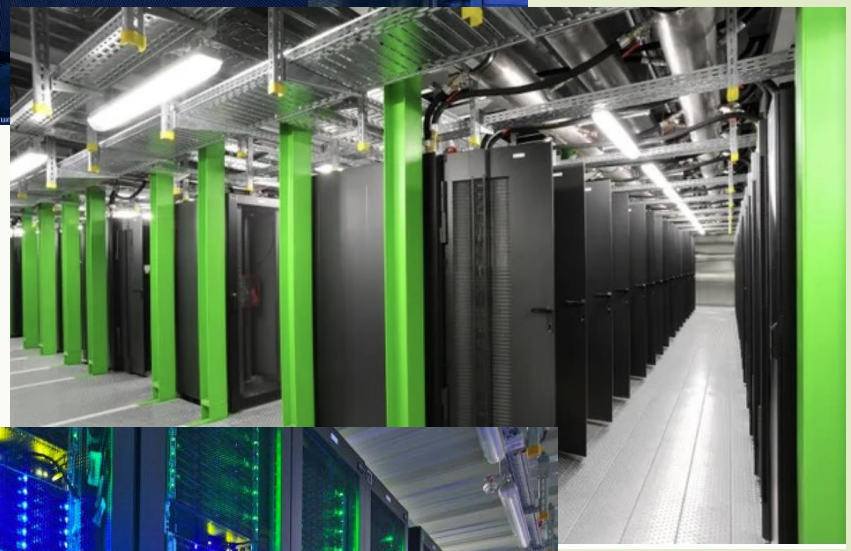
APPA

. . . . .

Computing status

- GC In operation since 2016
- Innovative cooling system (PUE: 1.07)
- Deployed 384 out of a total of 768 racks
- Total of 70.000 CPUs and 400 GPUs for computing
- 60 PB of storage
- Blue Angel energy efficiency certificate
- Provides also rackspace for external institutes and universities





FAIR: Computing Resource Requirements

	NUSTAR	CBM	PANDA	APPA	Theory, Bio, ..
Number of cores (a)	9k	45 k	68 k	11 k	~ 500 GPU
Number of cores (b)	7k	45 k	34 k	-	

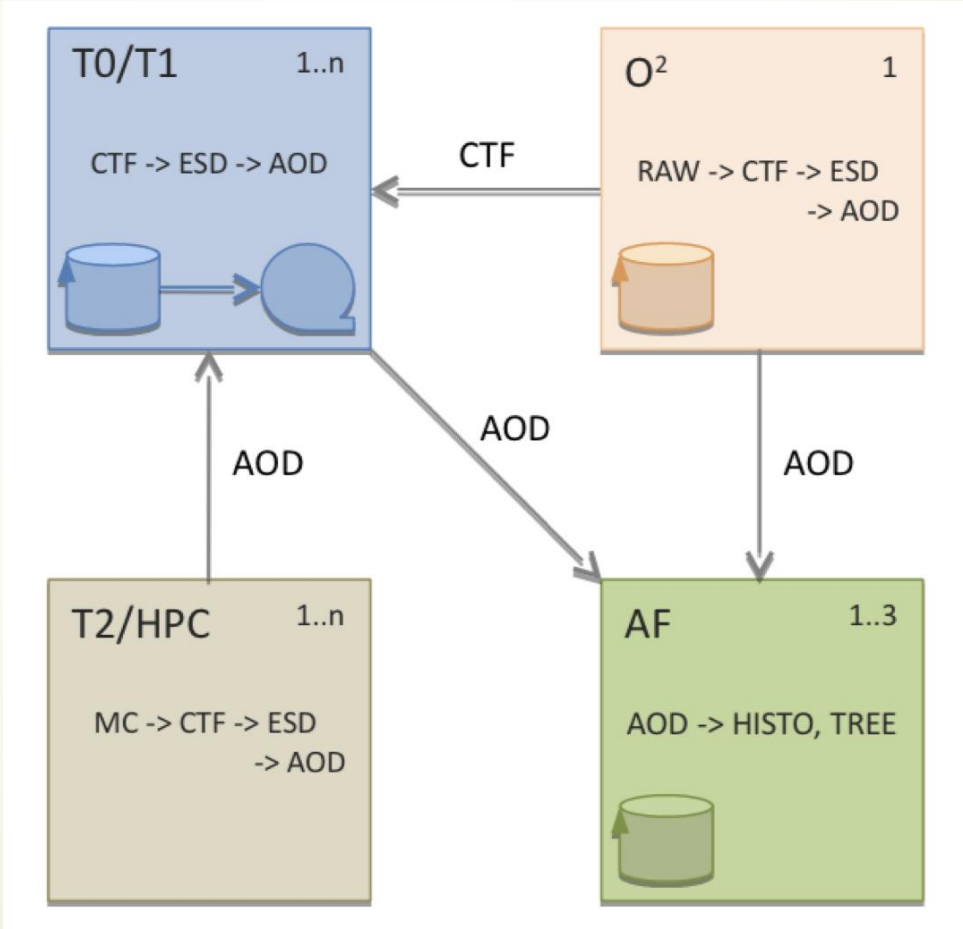
(a) Resources for simulations
(b) Resources for online data reconstruction

	NUSTAR	CBM	PANDA	APPA
Disk total (TB)	34.250	103.000	60.680	7.037

- No dedicated / fixed hardware for an experiment
- Will not take beam all at the same time
- Computing resources will be shared dynamically

The ALICE Analysis Facility at GSI

Analysis Facilities
(Grid Tier2)



AF Site resources

Analysis Facilities (Grid Tier2)

ALICE

- Resources reserved on a shared cluster:
- ~16k logical cores (hyperthreading) ⇔ 8k physical cores:

Nodes	CPUs/node	Total CPU	Memory/node
169	96	16224	192 GB

- 7 PiB disk storage under a Lustre distributed file system
- Network connection
  - internally 200 Gb/s HDR InfiniBand
  - 10 Gb/s LHCONE, 2 Gb/s DFN (research network)
- Memory limits imposed by Slurm via cgroups
  - limit is set on PSS ⇐ shared memory correctly accounted (4.4 GB per physical core)
  - no limit on virtual memory

GSI queues

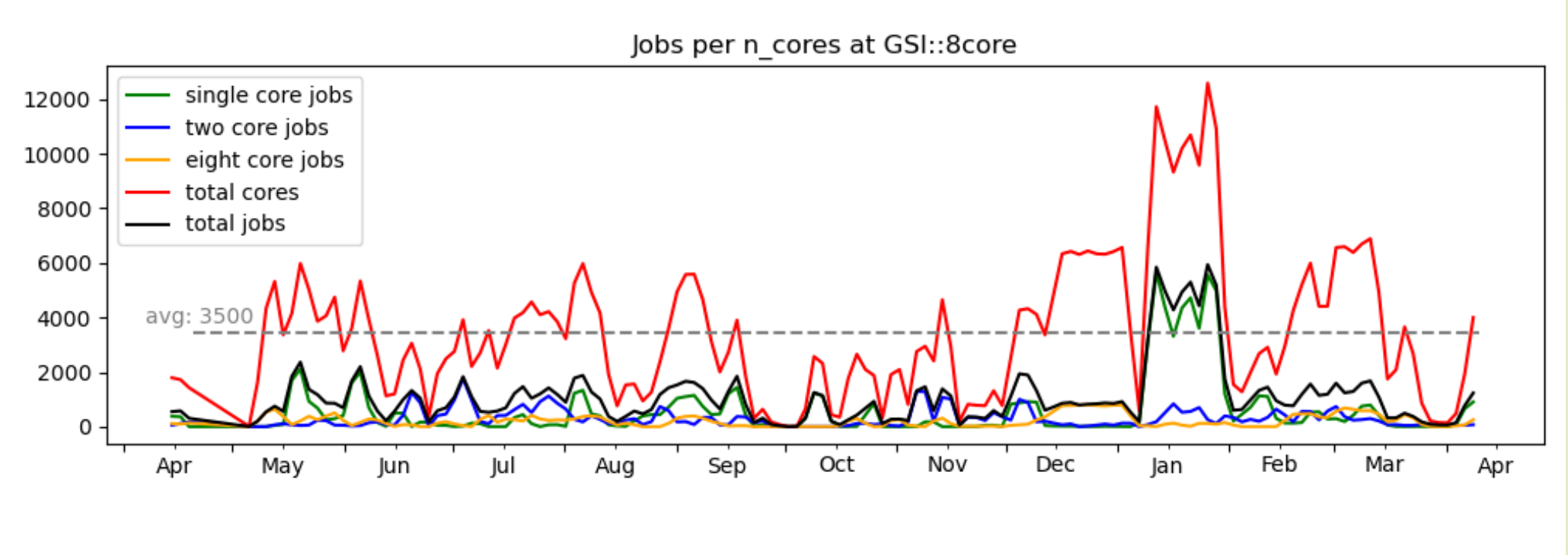
- Two queues at GSI: GSI_4core and GSI_8core.
  Since JobAgents started as Slurm 8-cores jobs get filled with single- and multi-core AliEn jobs, almost all jobs are queued on the 8-core VObox.
  Jobs run within Apptainer containers:
  - Host: minimal Red Hat Enterprise Linux compatible installation (Rocky 8.9)
  - Apptainer definition file and runtime engine managed by JAliEn (taken from /cvmfs/alice.cern.ch/containers/)

GSI queues

- In case of interest 16-core, high-memory and whole-node jobs are easy to provide as dedicated local queues with corresponding submit options

  (sbatch -c, --mem or --mem-per-cpu or --partition and --exclusive respectively).

Running jobs, core utilization

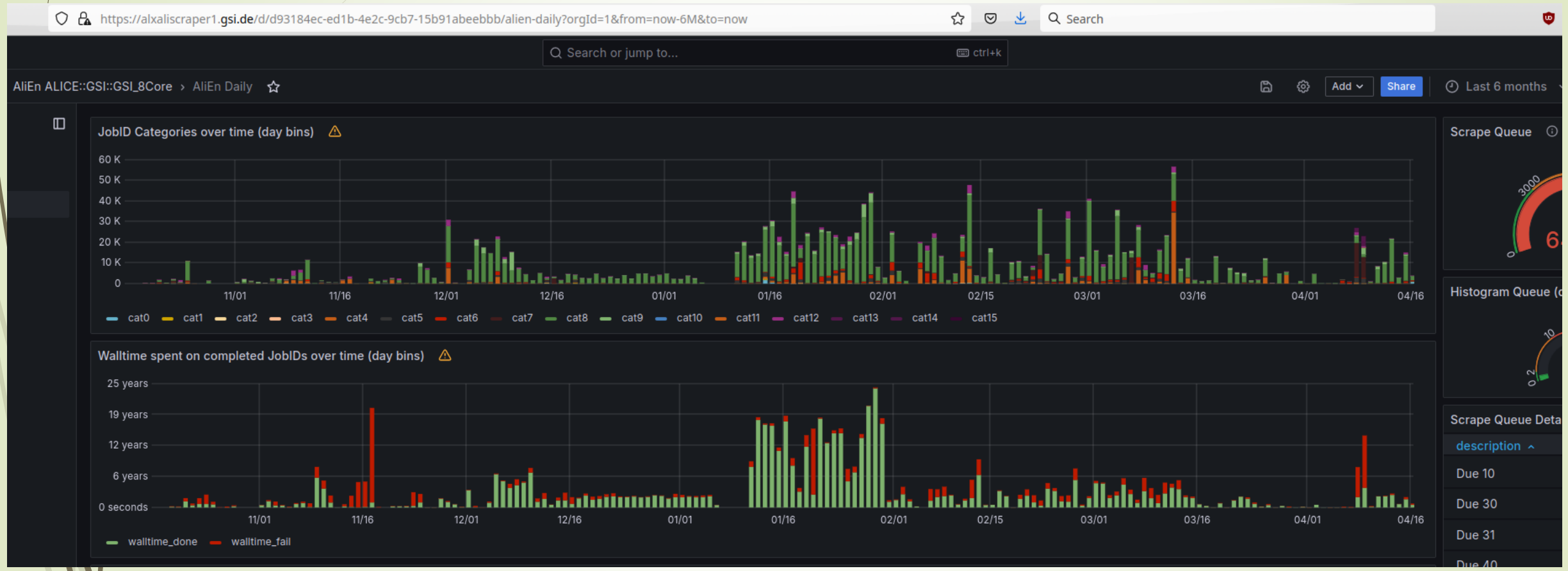


Average in the last year:

~3.5k utilized cores

~1.1k running jobs

Monitoring state of local jobs



Operations : new provisioning of servers

- Old bare metal machines have been decommissioned
- No more Debian
- 3 new(ish) xrootd data servers with Rocky Linux 8: alids{1-3}
- 2 xrootd redirectors (VMs) with Rocky Linux 8: alird{1-2}
- 1 NAT machine for outgoing ALICE traffic (+failover available on each xrootd data server)

Issues since last meeting

- CE fails silently (is running but cannot submit JobAgents). The PATH set in CE.env was not taken into account. Fixed in alienv. (Resolved)
- list of Apptainer binds not customizable, jobs cannot access local SE (/lustre) Fixed in jAliEn: additional directories to be bind mounted can be specified in the configuration. (Resolved)
- TPC transferred files get wrong permissions (0600 instead of 0644): it still requires a cronjob to chmod those files (should not concern transfers from client with xrootd version ≥5.5.0) (?)

Room for improvement

- In the past a good fraction of failed jobs at GSI were due to:
  - Failed to upload file due to: LFN already exists

Check in advance that multiple jobs don't try to upload the same LFN?

  - Out Of Memory

Use memory footprints from Hyperloop train tests?

e.g: for GSI queues: max(PSS) < 2.2GB × Ncores

Room for improvement

- Two local network issue were indirectly spotted out centrally looking at the CPU efficiency (CPU time/wall time) of Hyperloop trains at GSI
- we get now mail notifications for nightly trains to GSI and we have to check manually for train failures or low efficiencies! (hint to possible local issues)
- Could these checks be automated?

Anlysis Facilities in ALICE computing model

- AFs supposed to provide 50% of CPU share for analysis
- receive AODs from O2 farm and T1/T2s
- produce histograms and trees
- 10% of sampled AODs for quick analysis and cut tuning
- Requirements:
  - serve 6-8k job slots with ~15 MB/s/core2
  - aggregate throughput of 100 GB/s
  - be able to digest more than 5 PB of AODs in a 12-hour period

Analysis Facility at the GSI

- CPU resources are already available
- The required throughput (up to 100 GB/s) for worker nodes reading from the storage:
  - work ongoing to improve utilization of the shared Lustre file system
  - benchmarking of read throughput up to 6000 concurrent jobs planned

Plans for the near Future!

- Work ongoing to move to IPv4/IPv6 dual stack
- Dedicated storage instance