LHEP Site Report

Gianfranco Sciacca

AEC - Laboratory for High Energy Physics, University of Bern, Switzerland

HEPiX fall 2023 Victoria • 17 October 2023

UNIVERSITÄT
BERN

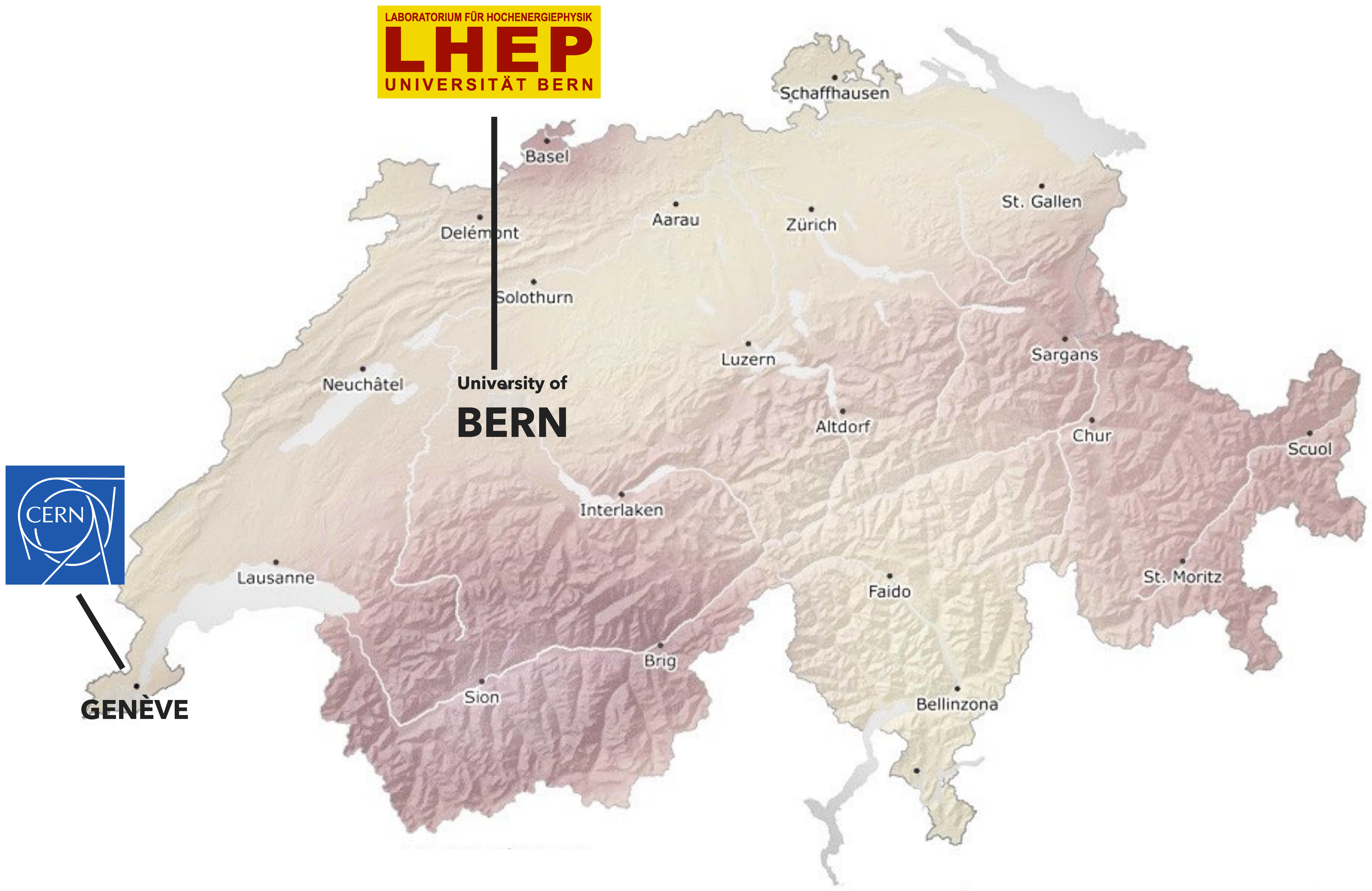
AEC
ALBERT EINSTEIN CENTER
FOR FUNDAMENTAL PHYSICS

LABORATORIUM FÜR HOCHENERGIEPHYSIK
LHEP
UNIVERSITÄT BERN

WHO AND WHERE

- The **Laboratory for High Energy Physics** is an institute of the Faculty of Science at the University of Bern, and also part of the Albert Einstein Centre for Fundamental Physics

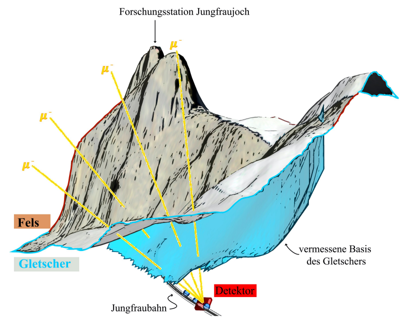
- **Research activities**:
  - High-Energy Collider Physics
  - Neutrino Physics
  - Fundamental Neutron and Precision Physics
  - Muon Radiography
  - Antimatter Physics
  - Development of Novel Particle Detectors
  - Medical Applications of Particle Physics

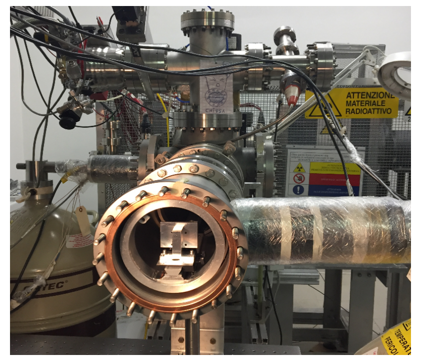


ATLAS EXPERIMENT

μBooNE

DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT


Forschungsstation Jungfraujoch
Fels
Gletscher
vermessene Basis des Gletschers
Jungfraubahn

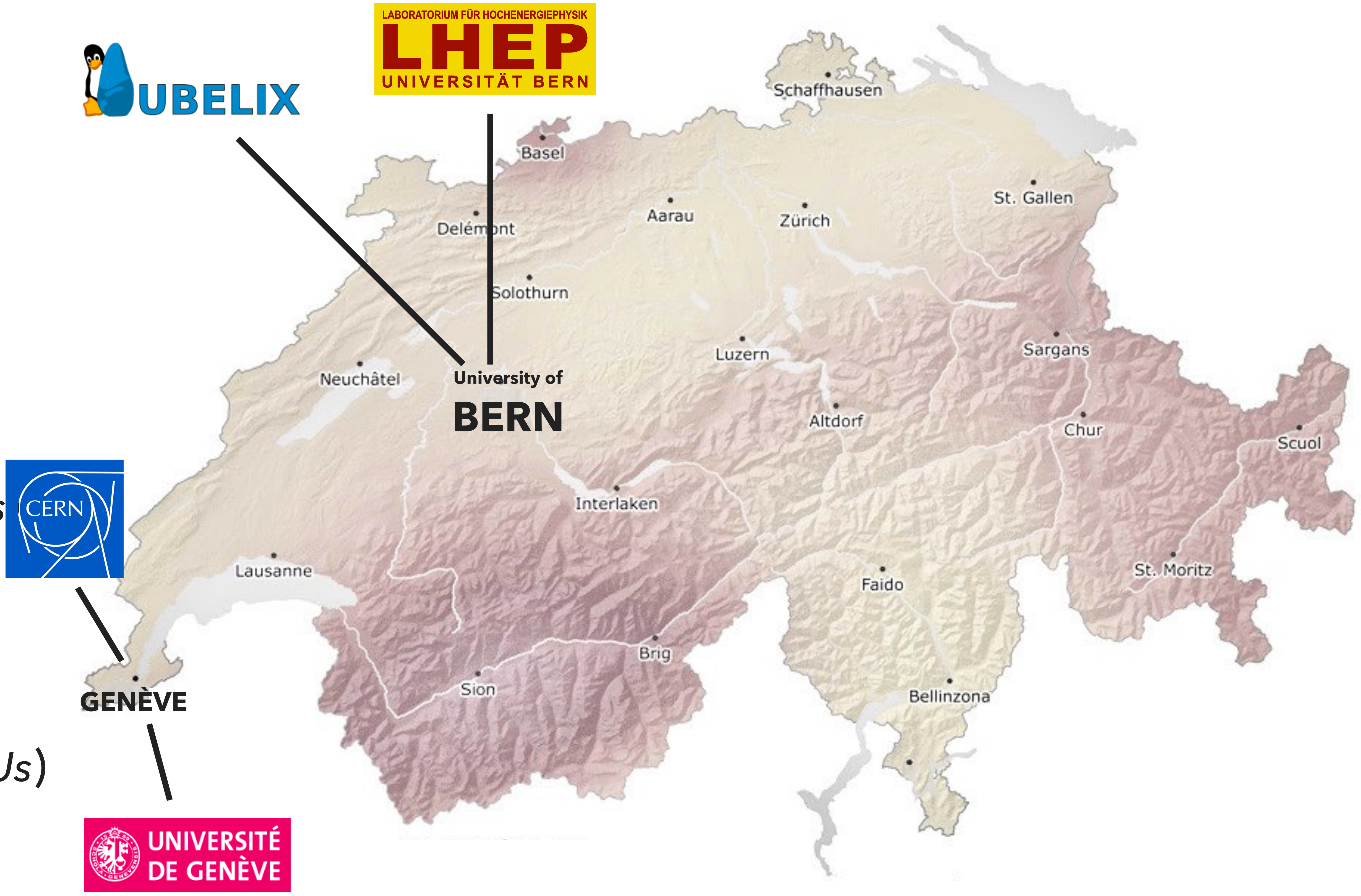


FASER

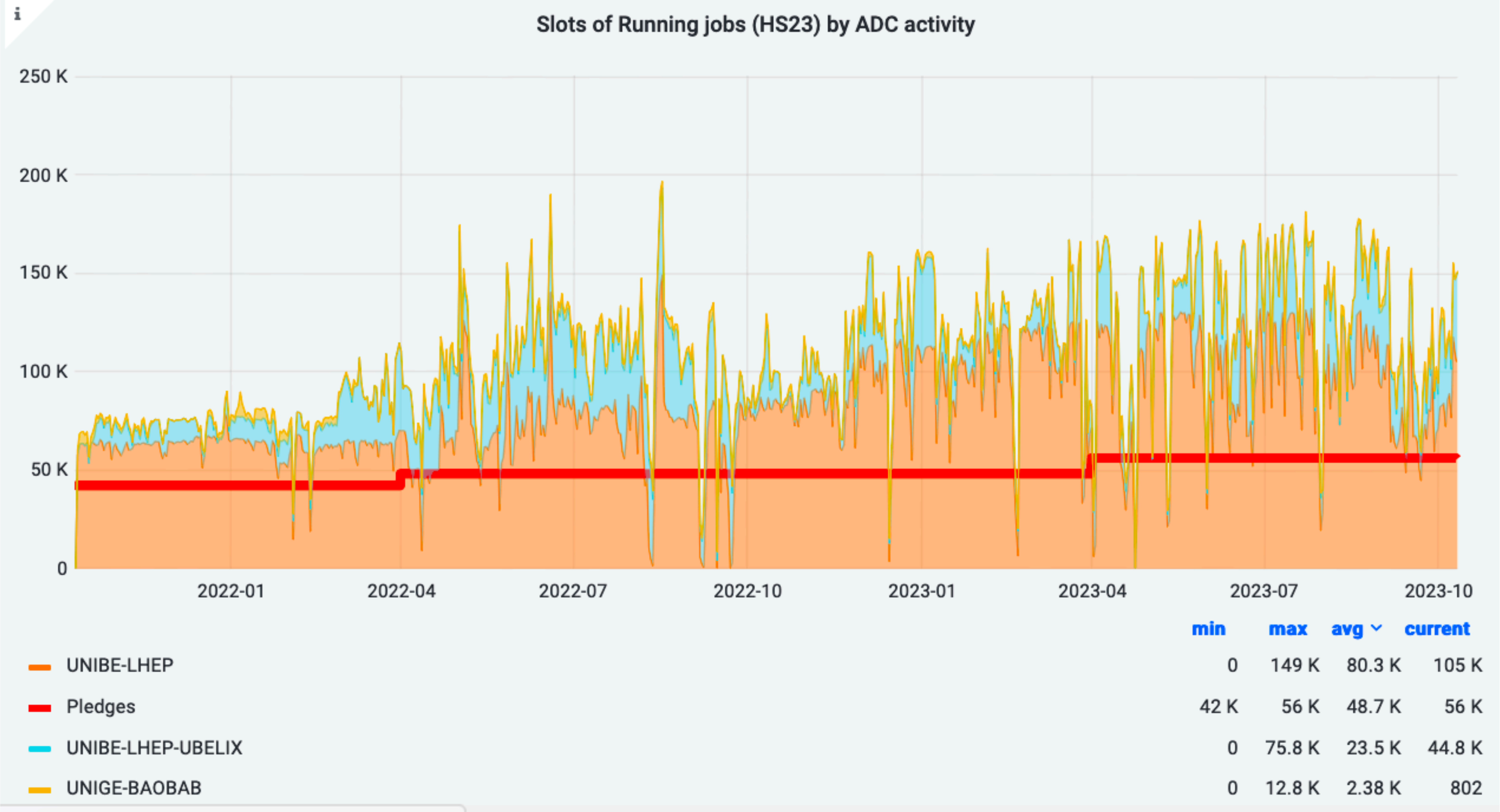
WHO AND WHERE

- **WCLG ATLAS Tier-2 since 2012**

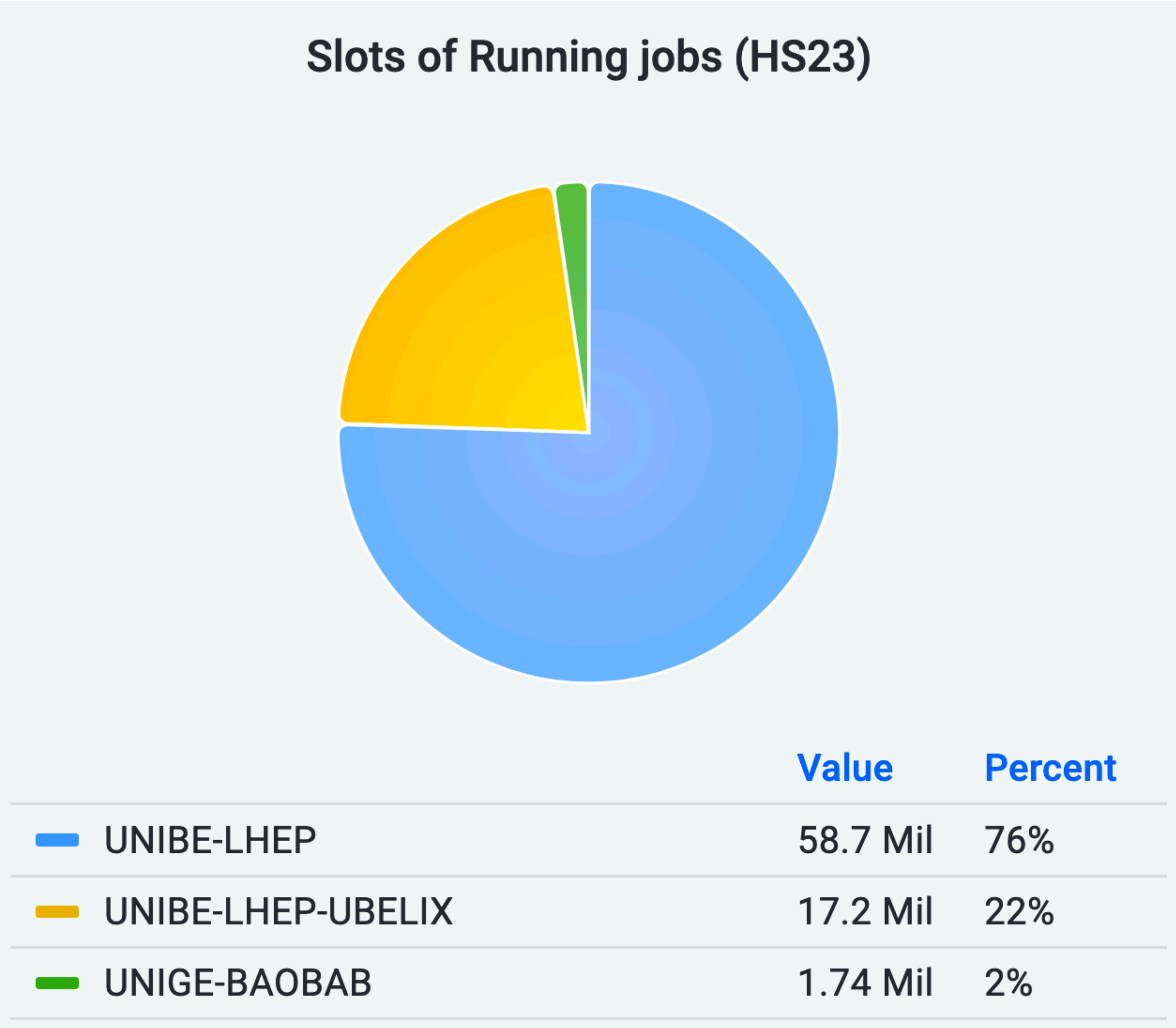
- **CH-ATLAS Federation @UniBE:**
  - ***LHEP*** dedicated resource:
    ~11k cores, Slurm, 0.5 PB ***Lustre*** cache+scratch,
    2.6 PB grid storage (*0.5 PB for neutrinos, also CPUs*)
  - ***UBELIX*** @UniBE (*multi-disciplinary cluster - 12k cores 160 GPUs*
    Slurm, ***up to 2k cores opportunistically***
    3.5 PB ***GPFS*** (for cache+scratch)
  - ***Baobab*** @UniGE (*multi-disciplinary cluster - 18k cores 320 GPUs*)
    Slurm, ***up tp 500 cores opportunistically***
    2.8 PB ***BeeGFS*** scratch
  - **Up to 180 kHS06 (*45 kHS06 opportunistic*)**



CPU IN THE ATLAS FEDERATION



	min	max	avg ⌄	current
UNIBE-LHEP	0	149 K	80.3 K	105 K
Pledges	42 K	56 K	48.7 K	56 K
UNIBE-LHEP-UBELIX	0	75.8 K	23.5 K	44.8 K
UNIGE-BAOBAB	0	12.8 K	2.38 K	802



	Value	Percent
UNIBE-LHEP	58.7 Mil	76%
UNIBE-LHEP-UBELIX	17.2 Mil	22%
UNIGE-BAOBAB	1.74 Mil	2%

1% of ATLAS Tier-2's in 2022

CPU IN THE ATLAS FEDERATION

LHEP

- AMD EPYC 7742 Rome + 4 Xeon generations
  1.3GB (80% of the cluster) to 4GB RAM per job slot

Ubelix

- AMD EPYC 7742 Rome
  4GB RAM per job slot

Baobab

- AMD EPYC 7742 Rome + Xeon E5-2630 v4
  3GB RAM per job slot

CPU IN THE ATLAS FEDERATION

LHEP

- AMD EPYC 7742 Rome + 4 Xeon generations
  1.3GB (80% of the cluster) to 4GB RAM per job slot

WORK IN PROGRESS

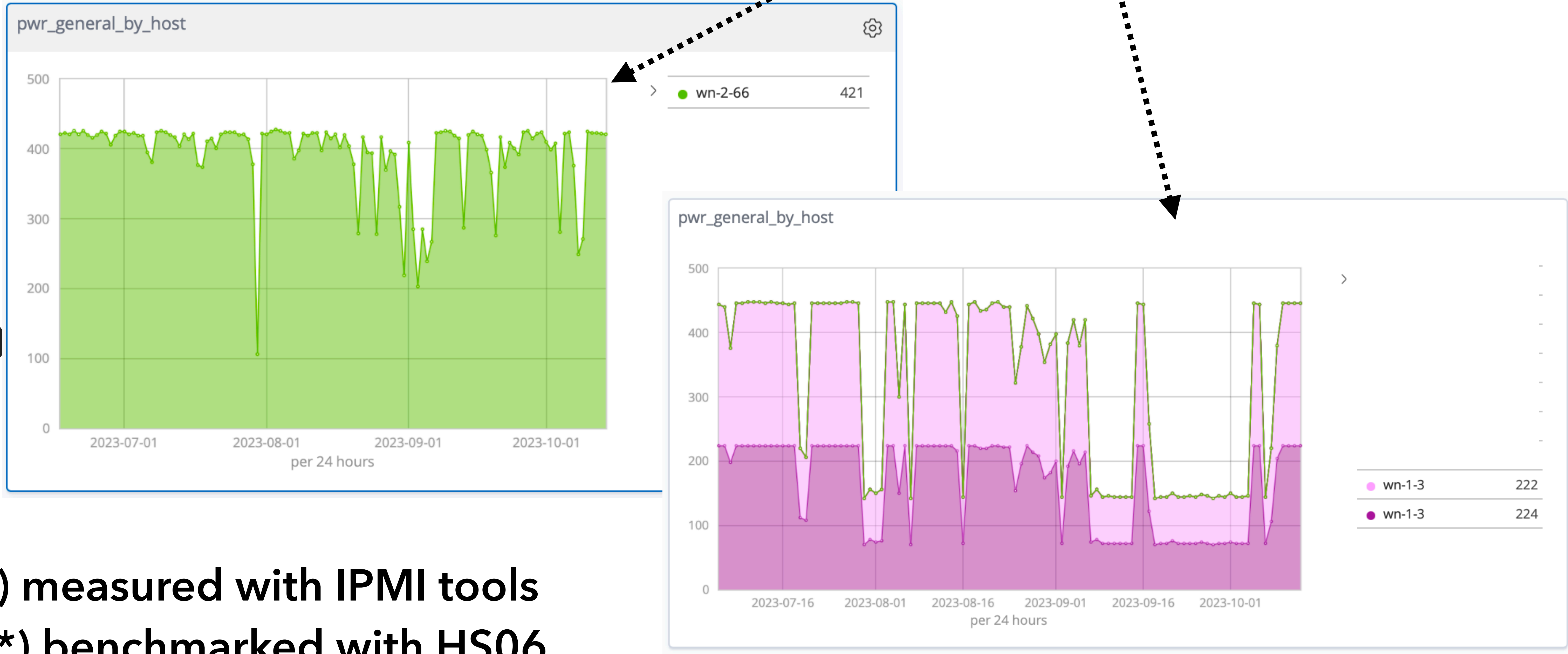
Node type @LHEP	Idle power	Full load power (*)	HS23/node (**)	HS23/Watt
AMD EPYC 7742 2x64-Core 2.25GHz HT on 512GB RAM	144W	450W	3158	7.02 ?
Intel Xeon E5-2680 v3 2x12-Core 2.50GHz HT on 64GB RAM	95W	420W	640	1.52

Ubelix

- AMD EPYC 7742 Rome
  4GB RAM per job slot

Baobab

- AMD EPYC 7742 Rome + Xeon E5-2630
  3GB RAM per job slot



(*) measured with IPMI tools
(**) benchmarked with HS06

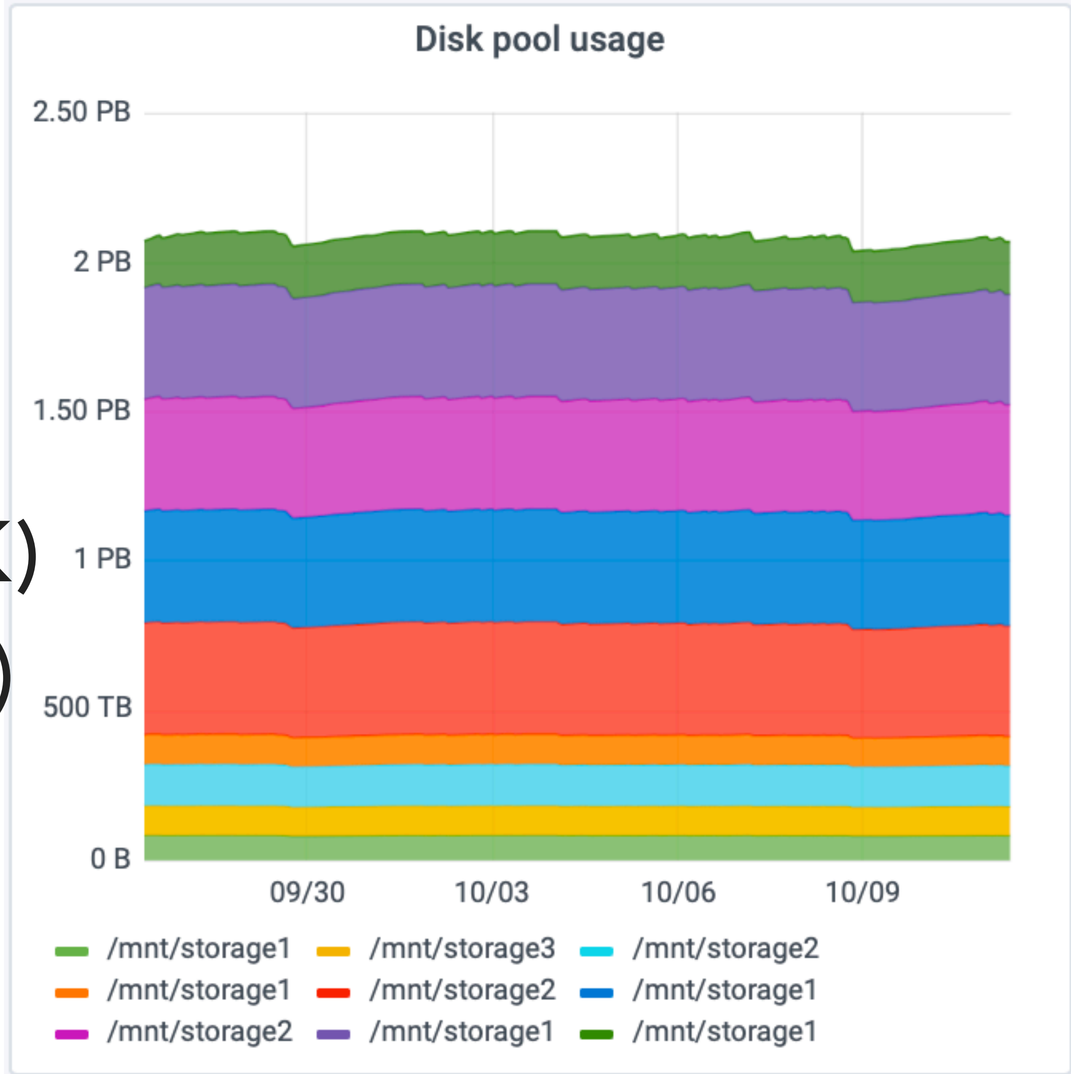
DISK STORAGE

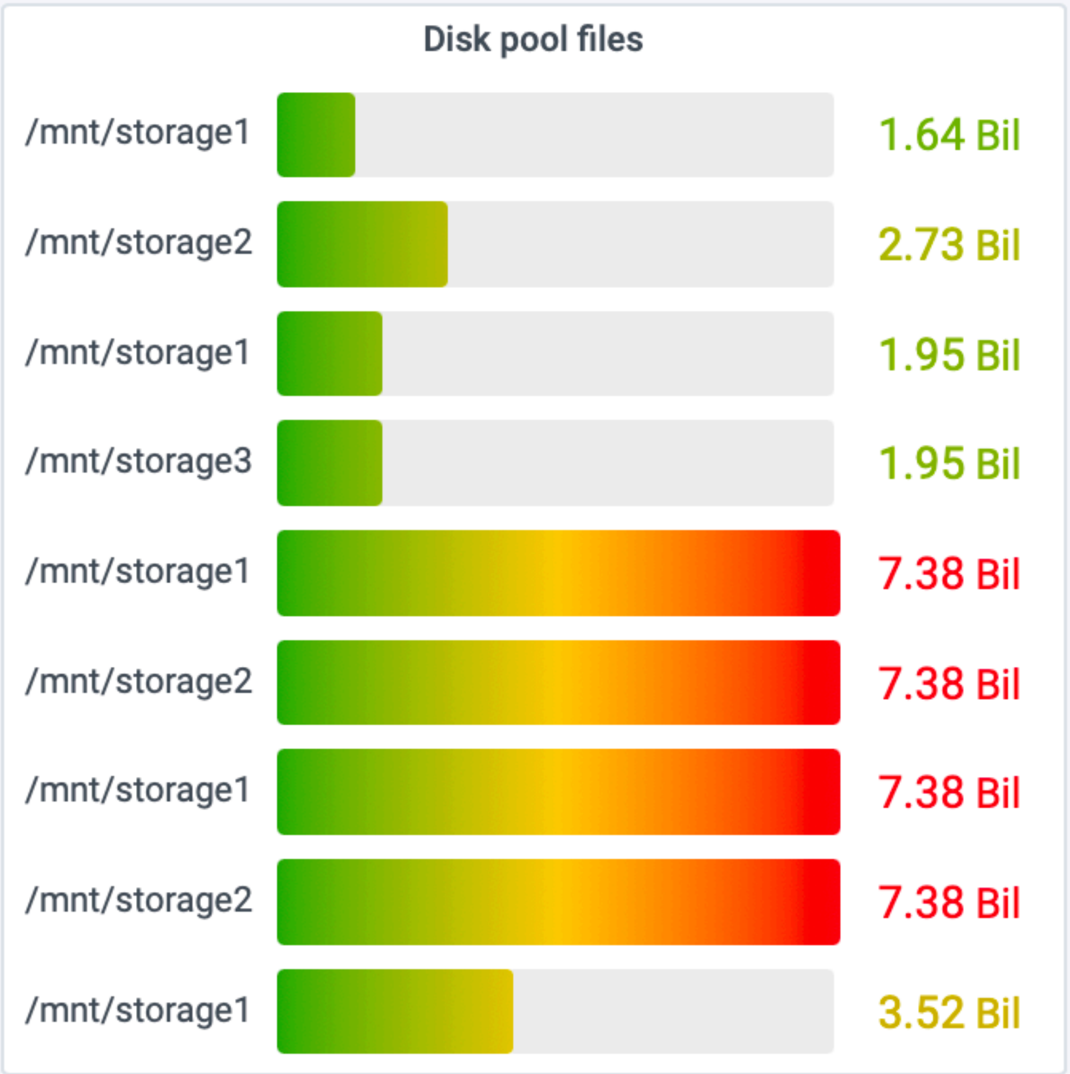
Grid storage @LHEP

A. 2.1 PB for ATLAS integrated with the NDGF-T1 dCache
   - RAID6 arrays, xfs, IPv4/6
   - 180 TB reservation for Swiss ATLAS users (LOCALGROUPDISK)
   - open issue with WLCG accounting (SRR for federated storage)

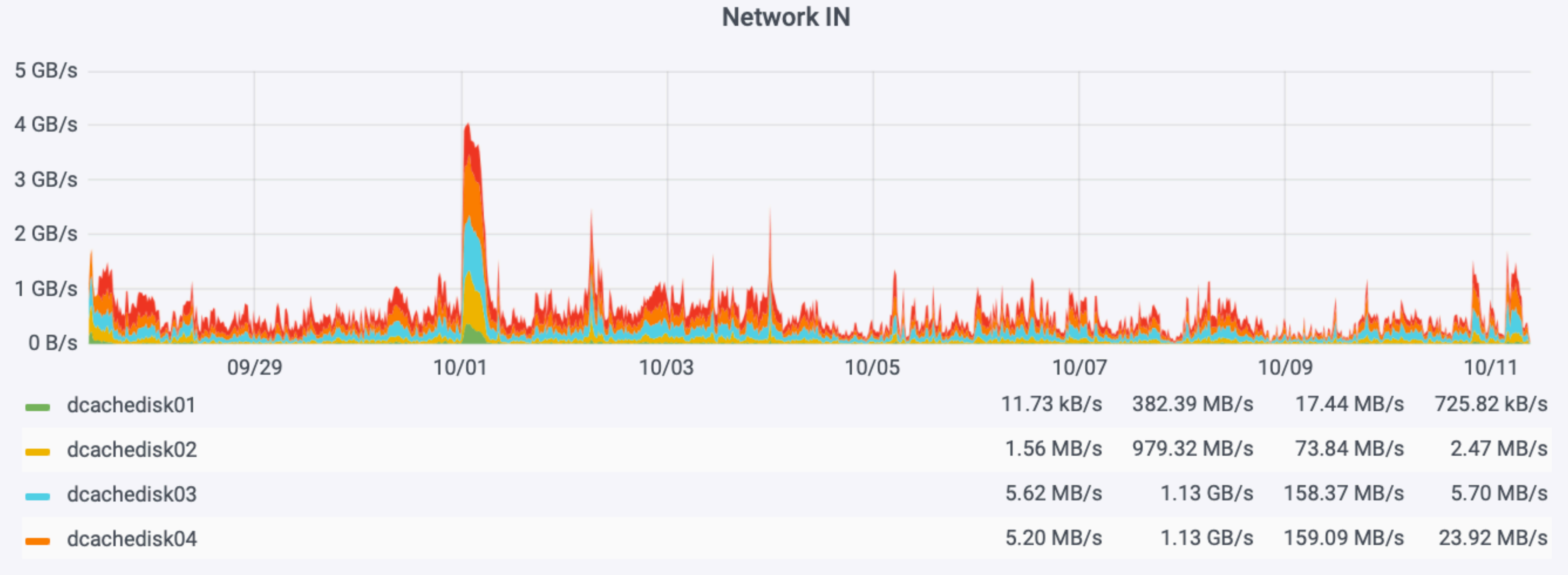
B. 0.5 PB for neutrinos in DPM, should migrate to dCache

Previous implementation: DPM federation Bern / Genève

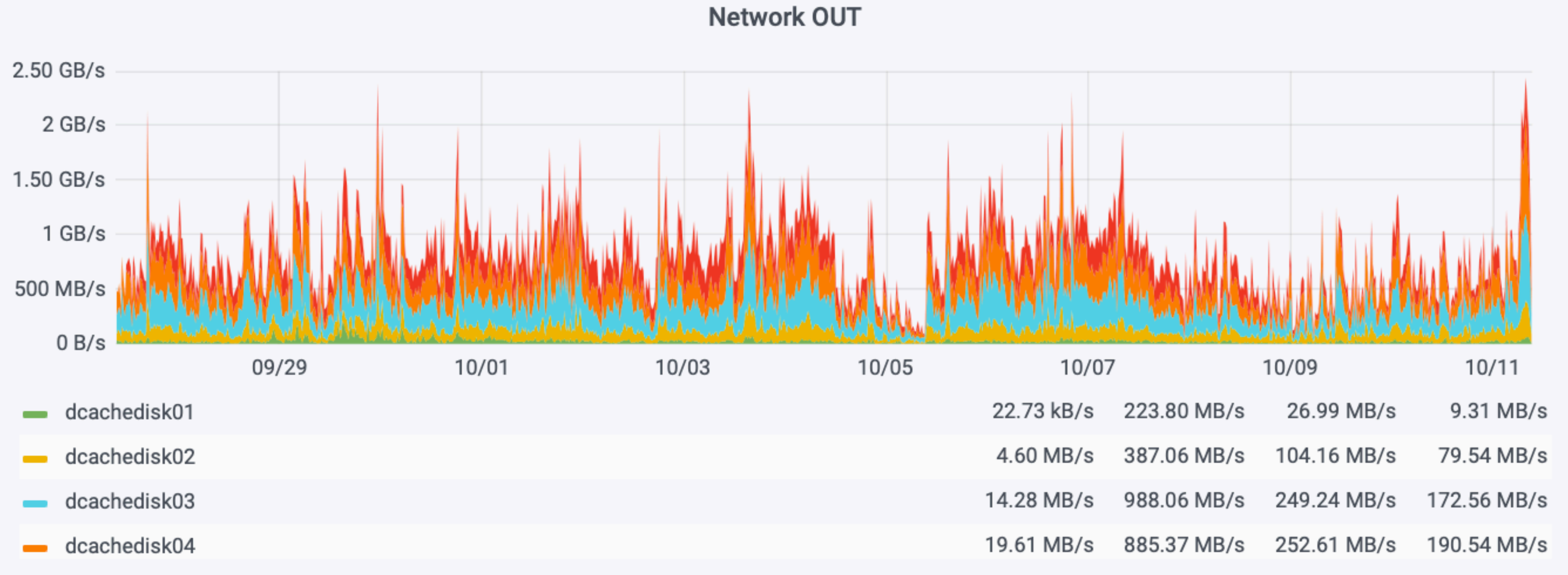



Disk pool files

/mnt/storage1	1.64 Bil
/mnt/storage2	2.73 Bil
/mnt/storage1	1.95 Bil
/mnt/storage3	1.95 Bil
/mnt/storage1	7.38 Bil
/mnt/storage2	7.38 Bil
/mnt/storage1	7.38 Bil
/mnt/storage2	7.38 Bil
/mnt/storage1	3.52 Bil



dcachedisk01	11.73 kB/s	382.39 MB/s	17.44 MB/s	725.82 kB/s
dcachedisk02	1.56 MB/s	979.32 MB/s	73.84 MB/s	2.47 MB/s
dcachedisk03	5.62 MB/s	1.13 GB/s	158.37 MB/s	5.70 MB/s
dcachedisk04	5.20 MB/s	1.13 GB/s	159.09 MB/s	23.92 MB/s



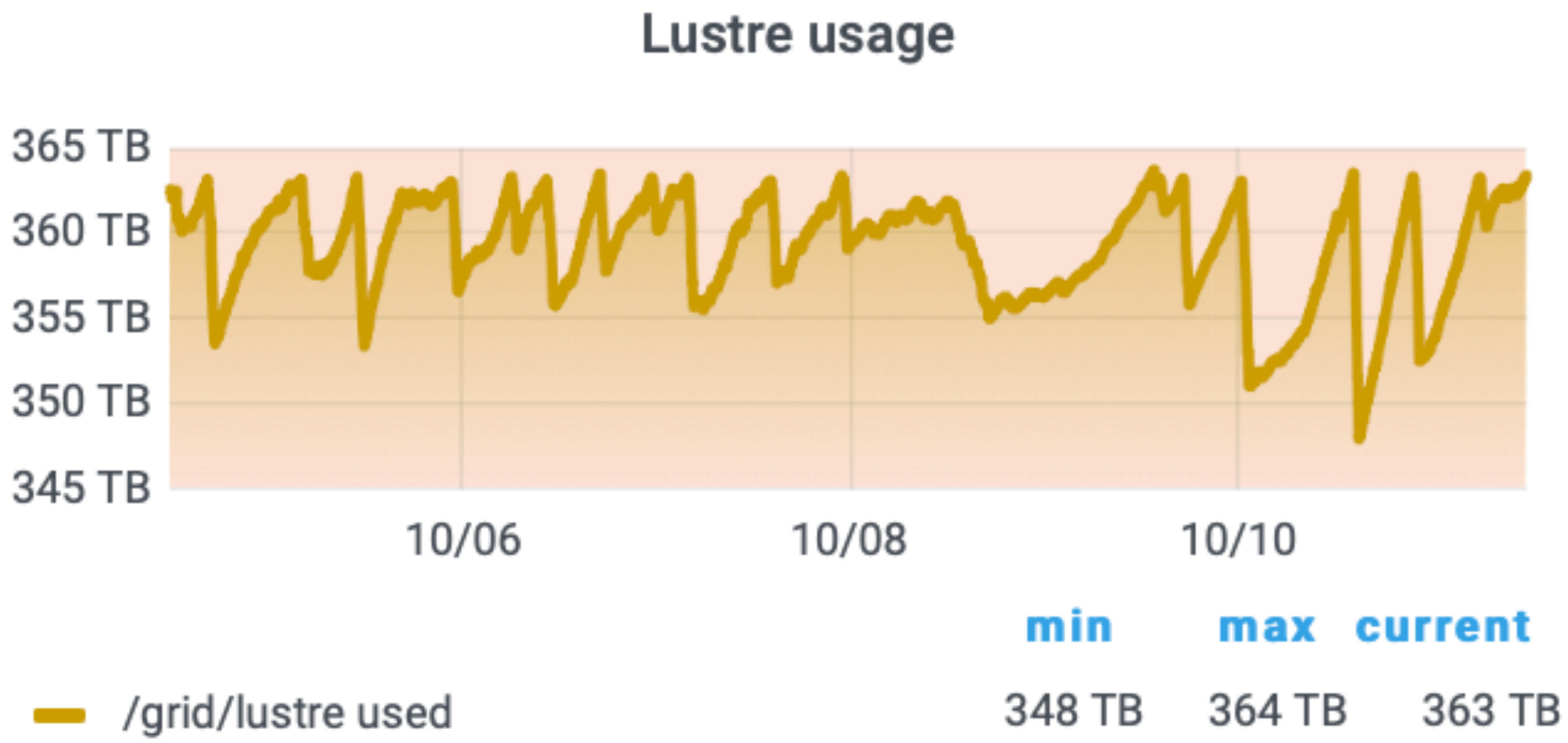
dcachedisk01	22.73 kB/s	223.80 MB/s	26.99 MB/s	9.31 MB/s
dcachedisk02	4.60 MB/s	387.06 MB/s	104.16 MB/s	79.54 MB/s
dcachedisk03	14.28 MB/s	988.06 MB/s	249.24 MB/s	172.56 MB/s
dcachedisk04	19.61 MB/s	885.37 MB/s	252.61 MB/s	190.54 MB/s

DISK STORAGE

Other storage @LHEP

A. Cluster

- 500 TB in Lustre for ARC cache (low latency data access) and job scratch areas
  - mdadm arrays for OSTs, HDDs
  - fs usage profile dominated by the cache cleaning routine



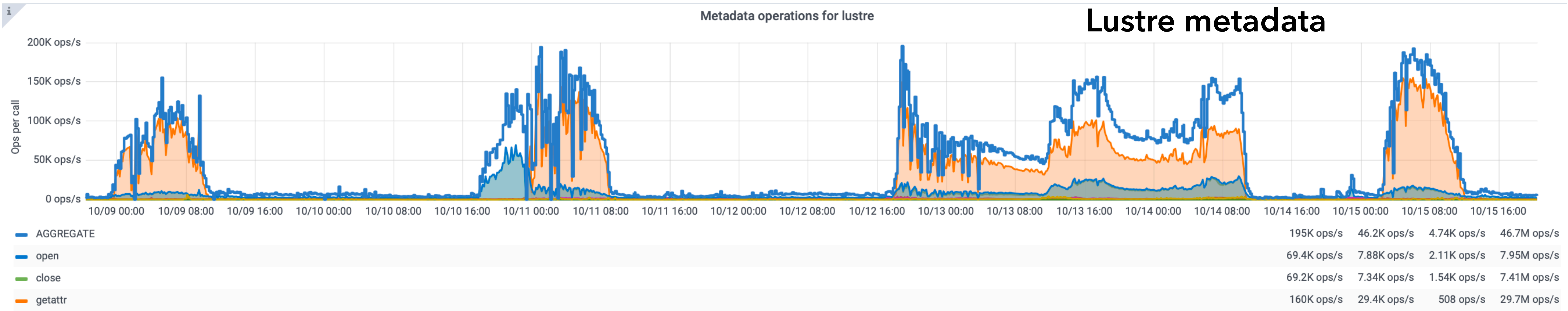
B. Interactive and labs

- 100 TB in Ceph for interactive local users
  - grid is preferred by ATLAS users (e.g. LOCALGROUPDISK), but some local storage needed for a few applications
- NFS for home directories on the interactive platform (with backup)
- Scattered storage for other users, labs, not centrally managed (typically NFS, also NAS appliances)
  - moving towards better integration on shared resources

DISK STORAGE

Lustre @LHEP

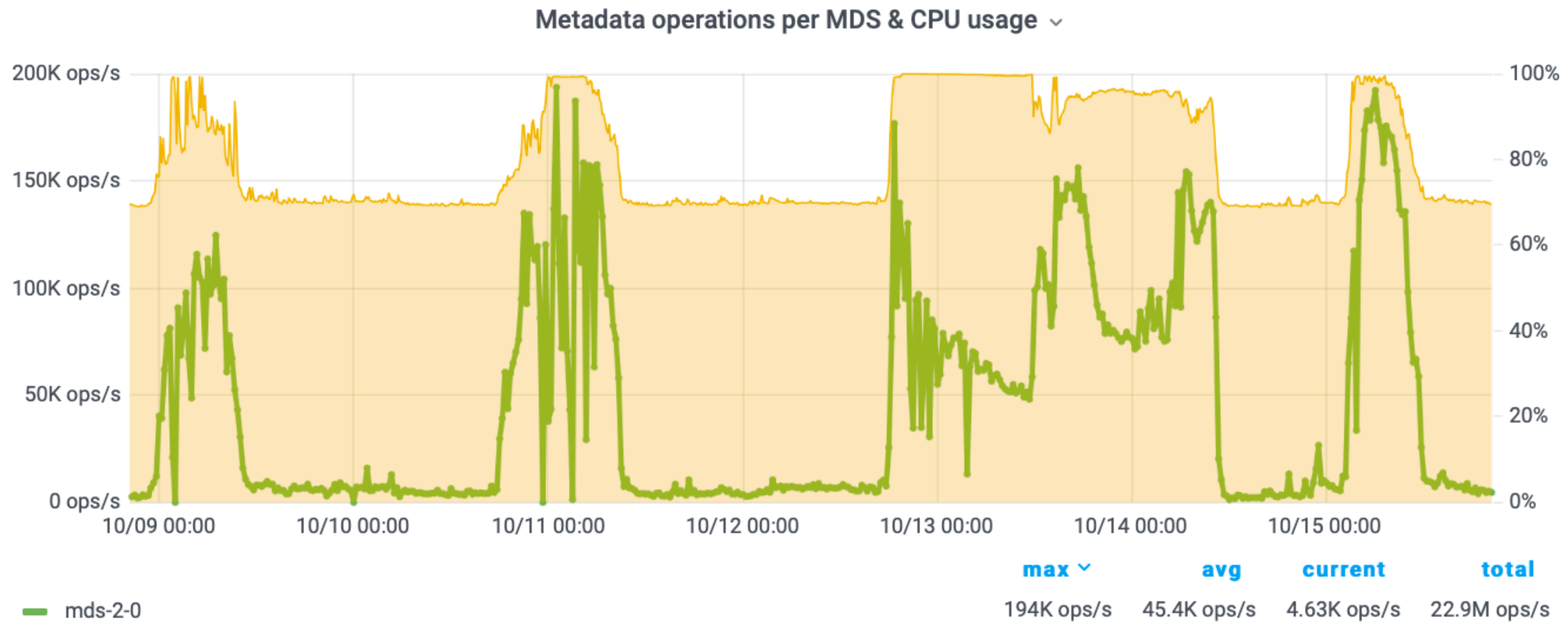
- version 2.12.9-1



Metadata operations for lustre

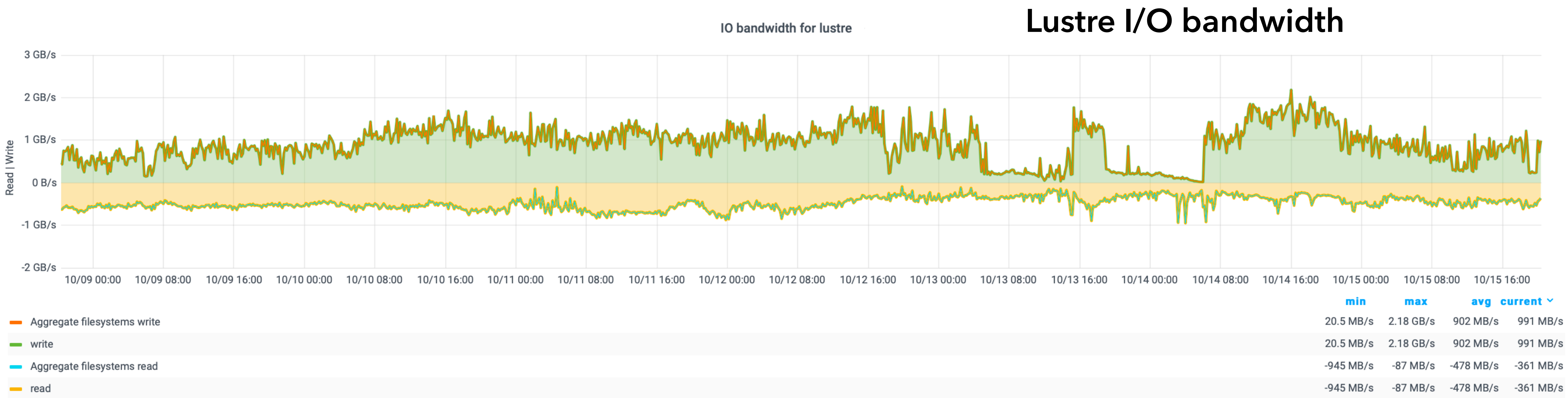
Lustre metadata

AGGREGATE	195K ops/s	46.2K ops/s	4.74K ops/s	46.7M ops/s
open	69.4K ops/s	7.88K ops/s	2.11K ops/s	7.95M ops/s
close	69.2K ops/s	7.34K ops/s	1.54K ops/s	7.41M ops/s
getattr	160K ops/s	29.4K ops/s	508 ops/s	29.7M ops/s



Metadata operations per MDS & CPU usage

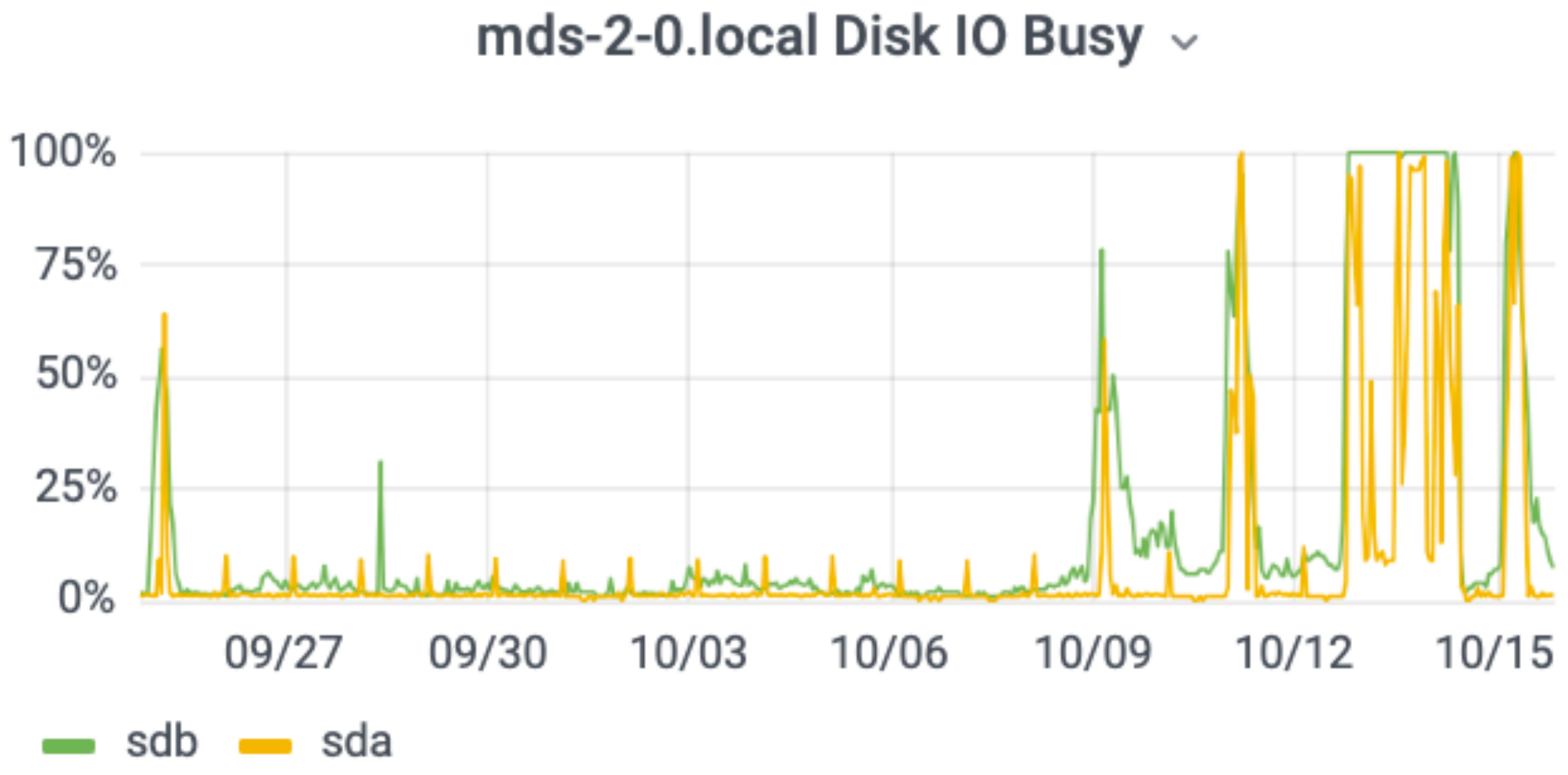
	max	avg	current	total
mds-2-0	194K ops/s	45.4K ops/s	4.63K ops/s	22.9M ops/s



IO bandwidth for lustre

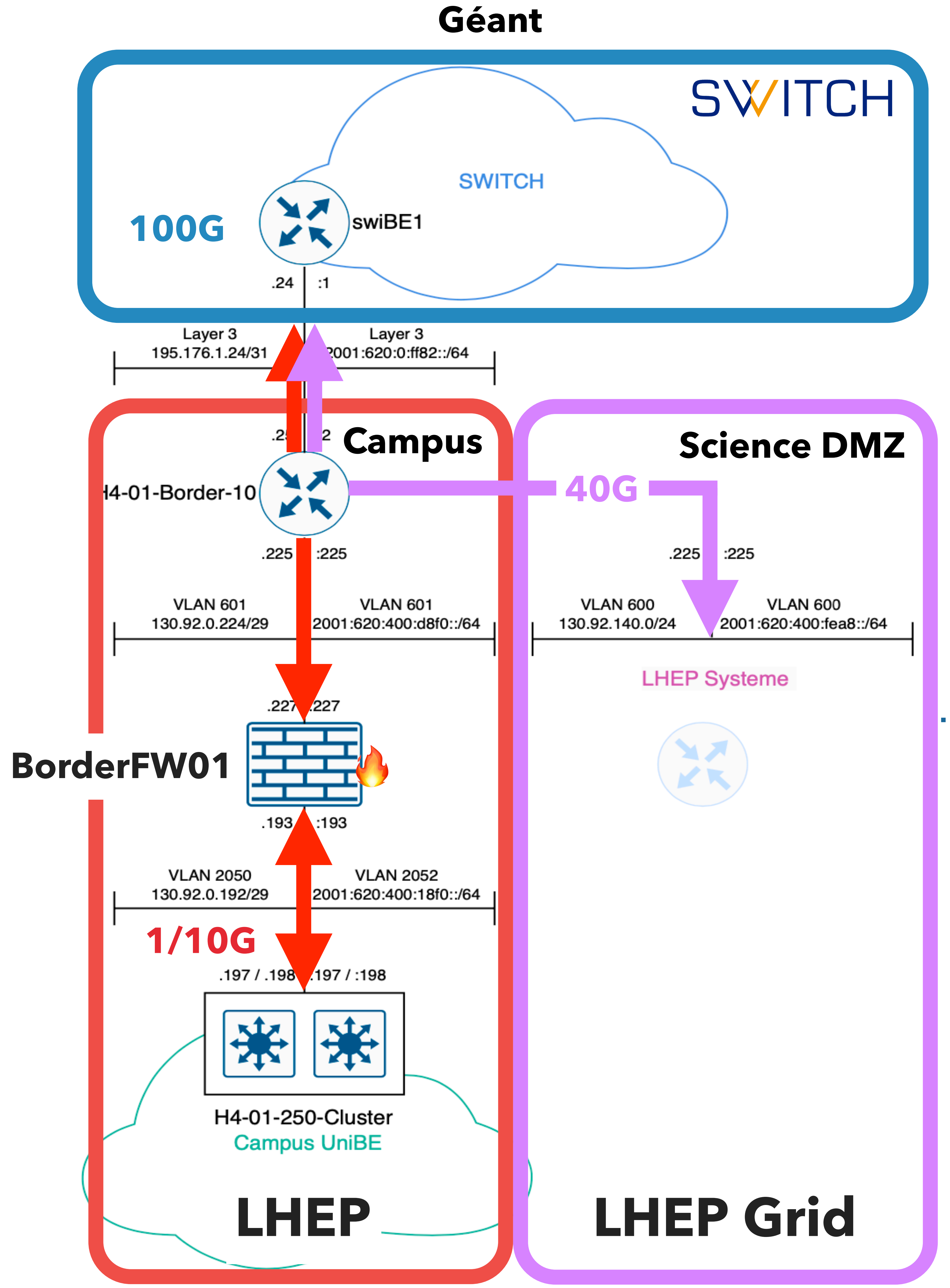
Lustre I/O bandwidth

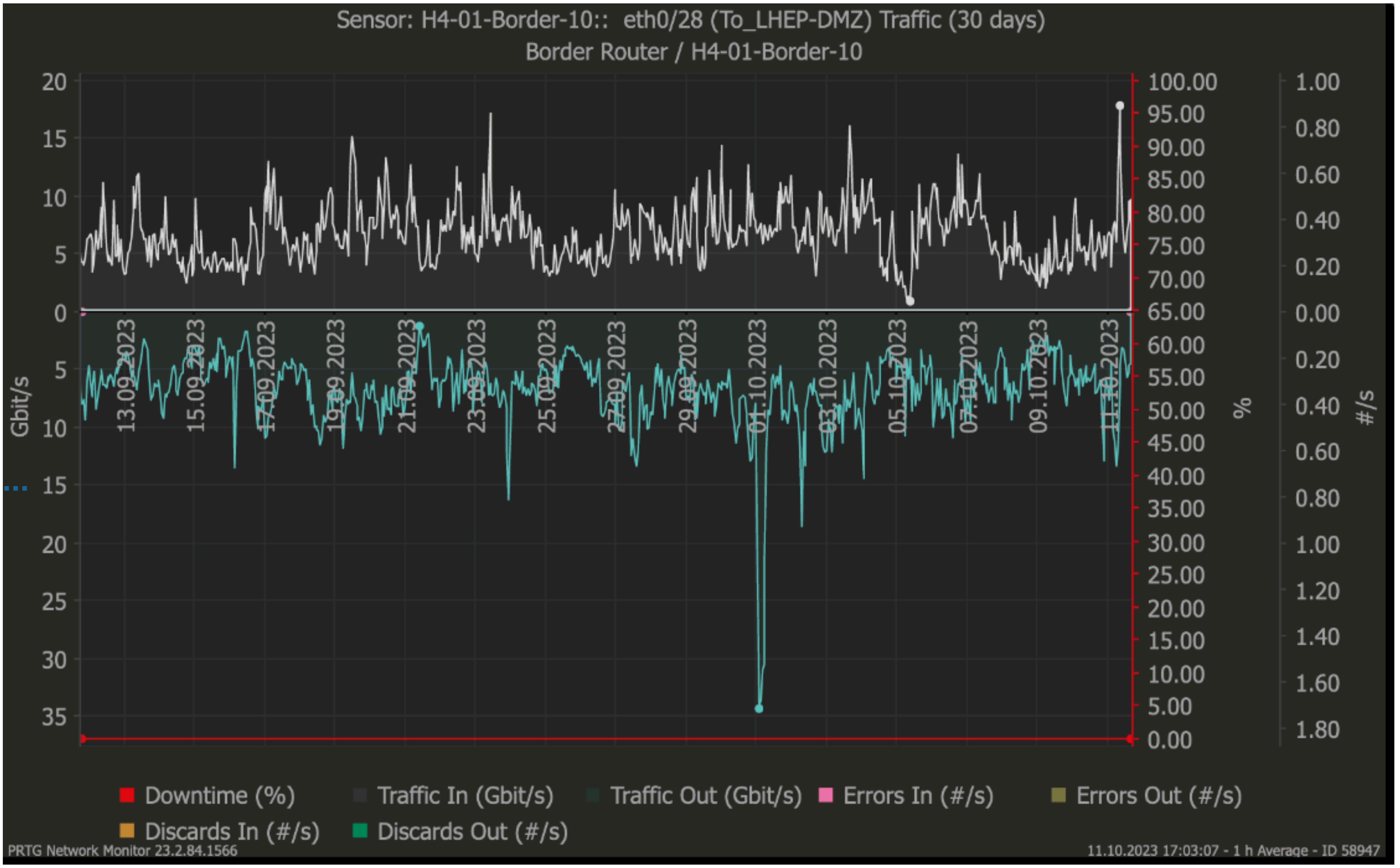
	min	max	avg	current
Aggregate filesystems write	20.5 MB/s	2.18 GB/s	902 MB/s	991 MB/s
write	20.5 MB/s	2.18 GB/s	902 MB/s	991 MB/s
Aggregate filesystems read	-945 MB/s	-87 MB/s	-478 MB/s	-361 MB/s
read	-945 MB/s	-87 MB/s	-478 MB/s	-361 MB/s



mds-2-0.local Disk IO Busy

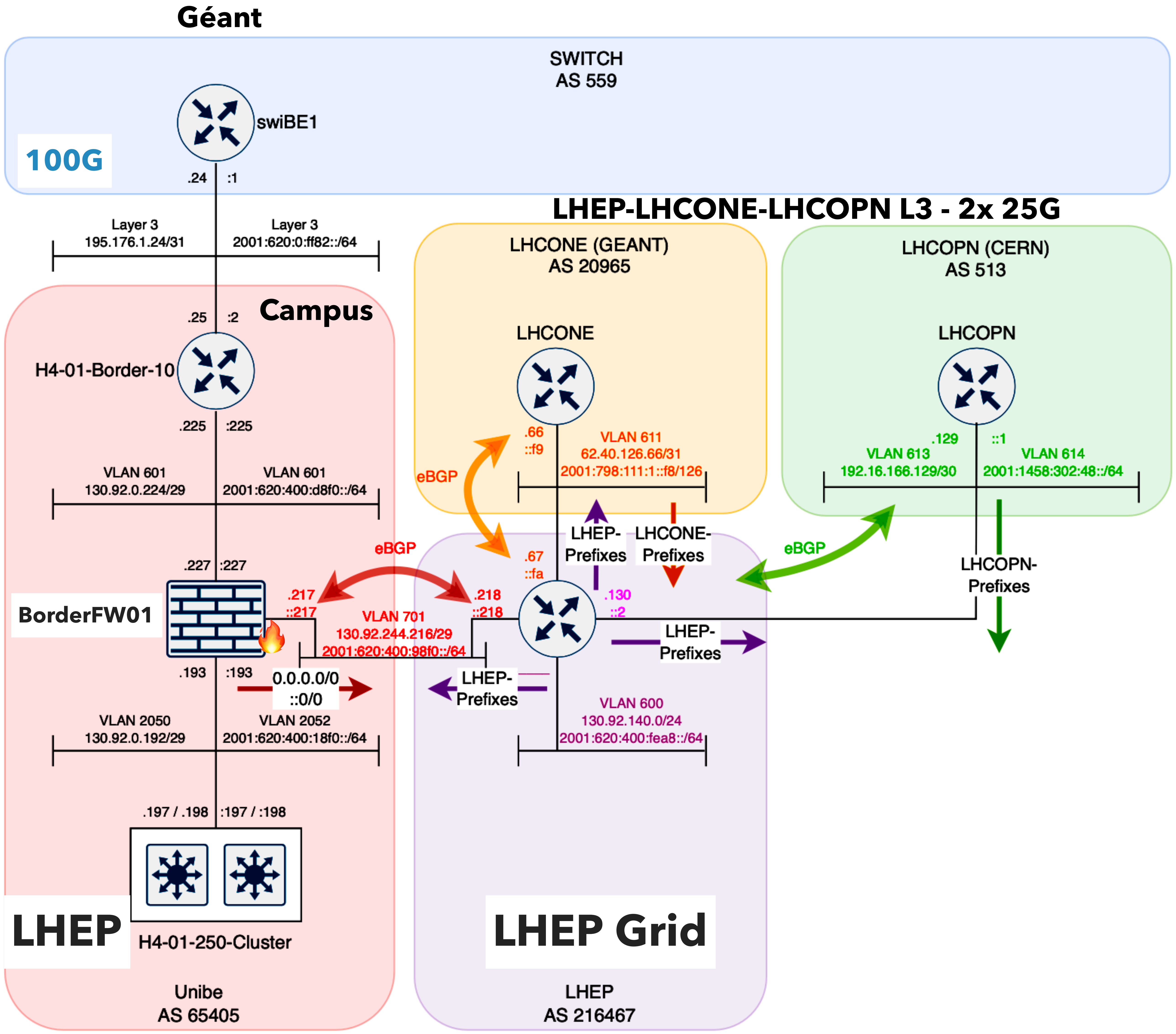
NETWORK



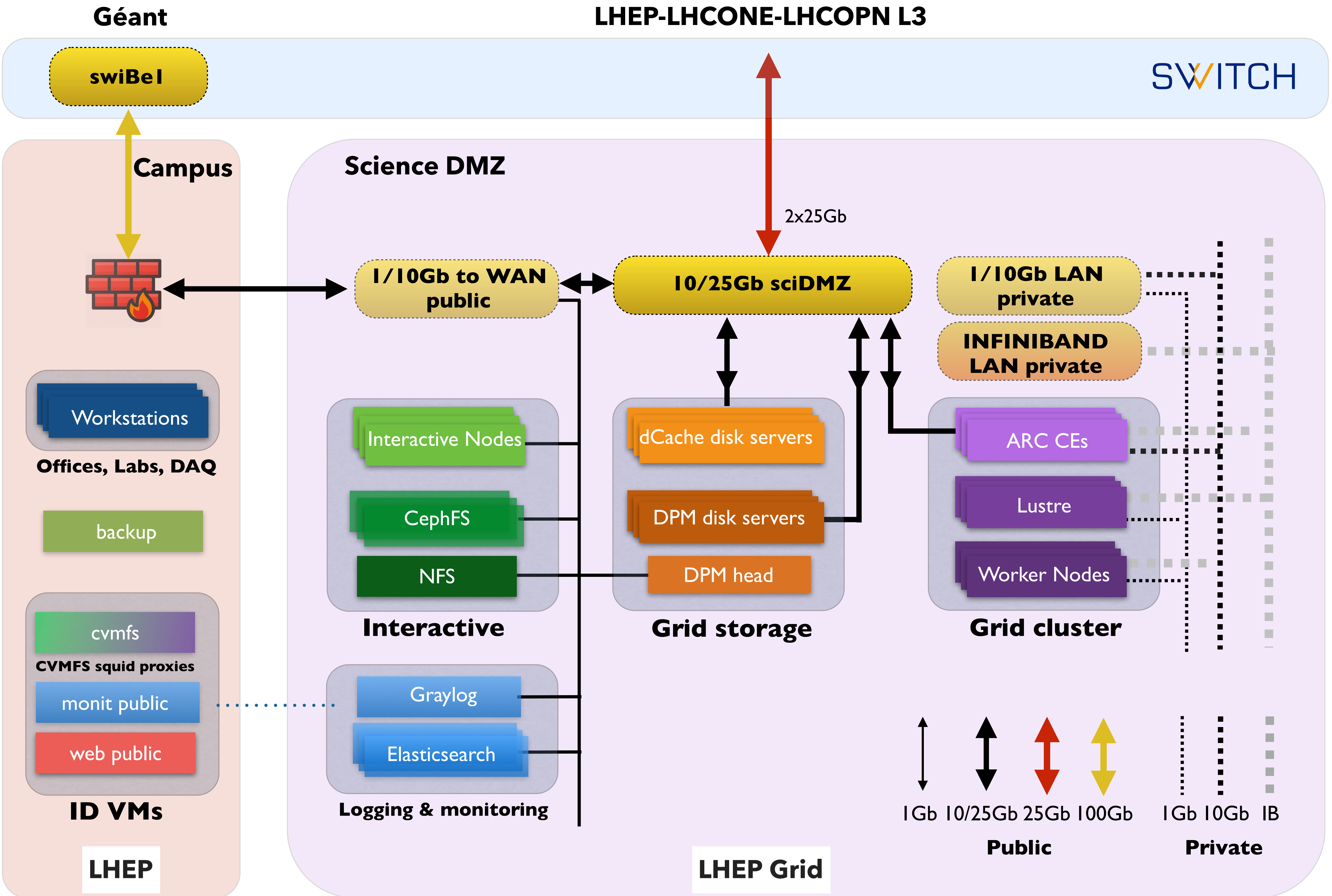

Sensor: H4-01-Border-10:: eth0/28 (To_LHEP-DMZ) Traffic (30 days)
Border Router / H4-01-Border-10

NETWORK (NEW!)

TRANSITION PLANNED FOR 30 NOVEMBER 2023



SERVICES @LHEP



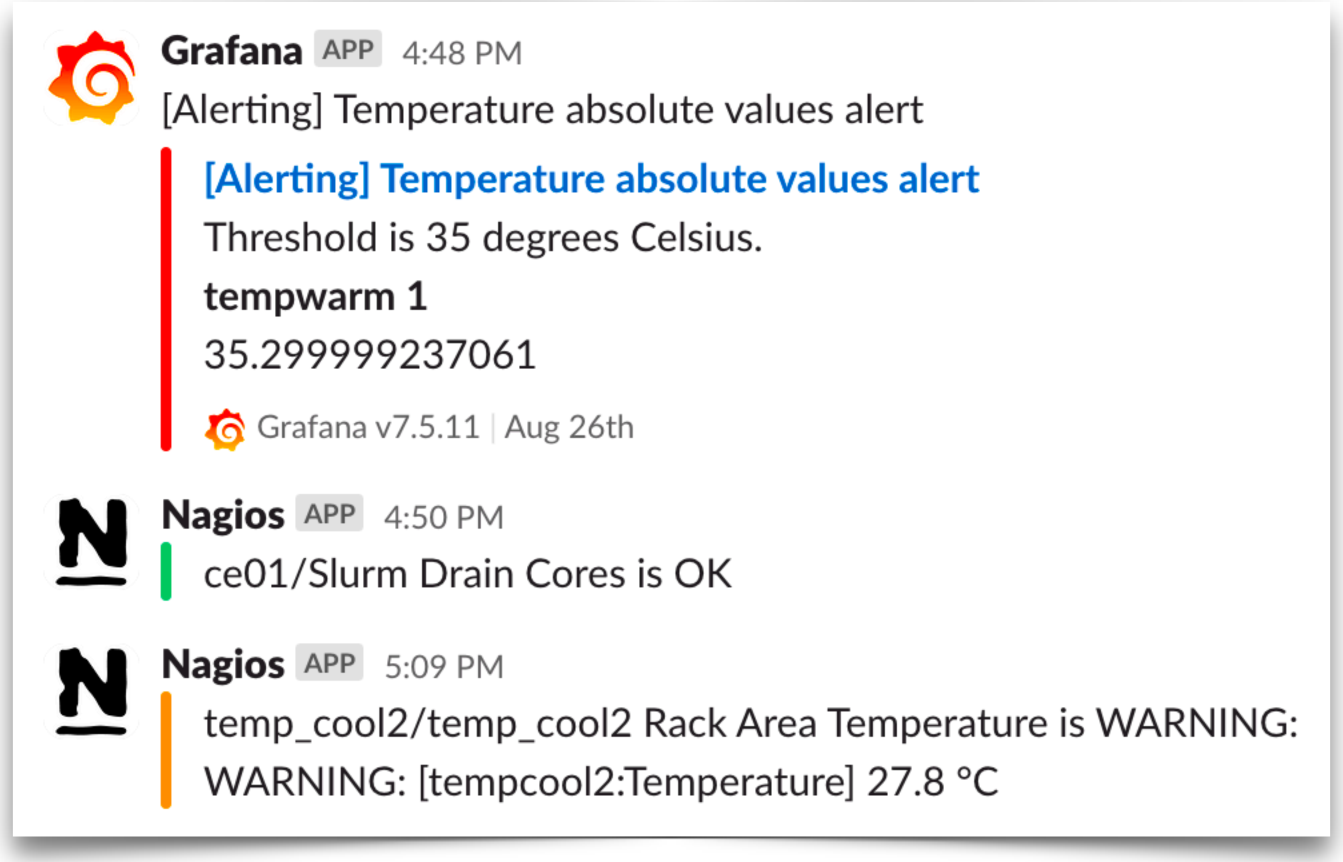
LOGGING / MONITORING / ALERTING

Graylog + Beats + Elasticsearch + Kibana + Grafana

- **syslog** from all managed servers redirected to a central **graylog** instance
- **metricbeat** ships system metrics to **graylog**
  - from cluster directly to **elasticsearch** (phasing out **ganglia**)
- **filebeat** for custom data collection (e.g. ipmi)
- **heartbeat** (uptime, http/s), **auditbeat** (security)
- **prometheus** (slurm, lustre, infiniband, ARC)
- **elasticsearch** backend, small 5-node data cluster
  - 1-year log retention
  - 30-day retention for metricbeat data
  - 6-12 months retention for other beats / prometheus data
- **kibana** and **grafana** for visualisation

Nagios

- alerting (email+slack) for all managed resources
- a few alerts for mission critical metrics duplicated in grafana

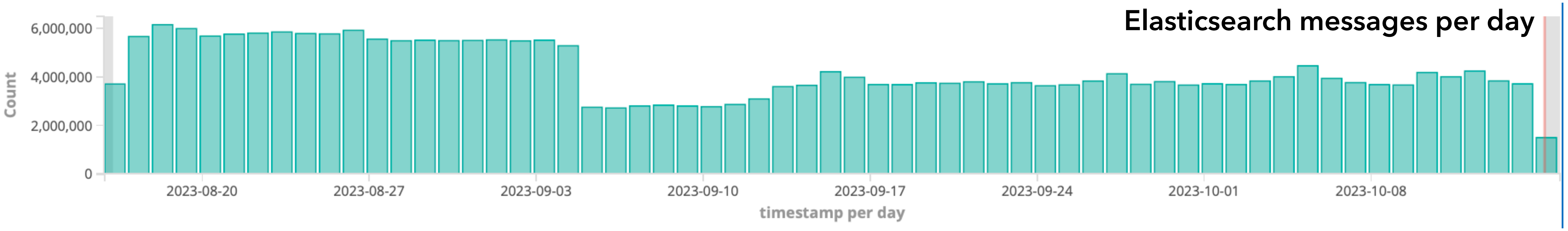


Grafana APP 4:48 PM
[Alerting] Temperature absolute values alert

[Alerting] Temperature absolute values alert
Threshold is 35 degrees Celsius.
tempwarm 1
35.299999237061
Grafana v7.5.11 | Aug 26th

Nagios APP 4:50 PM
ce01/Slurm Drain Cores is OK

Nagios APP 5:09 PM
temp_cool2/temp_cool2 Rack Area Temperature is WARNING:
WARNING: [tempcool2:Temperature] 27.8 °C



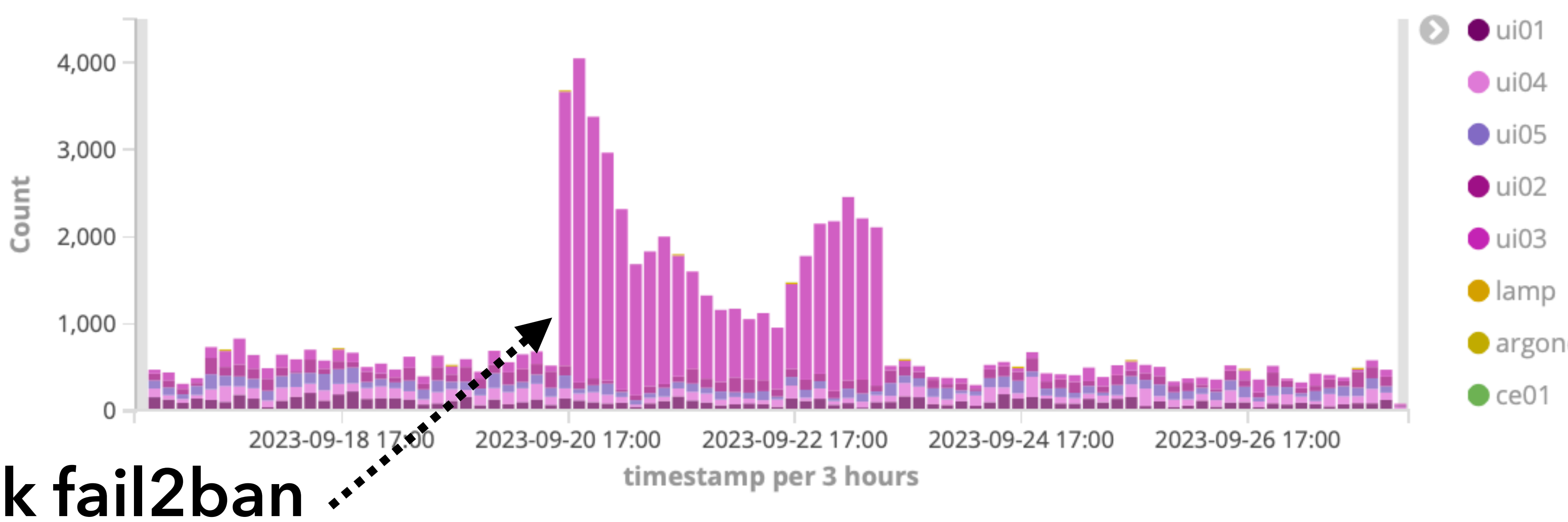
Elasticsearch messages per day

LOGGING / MONITORING / ALERTING

Security

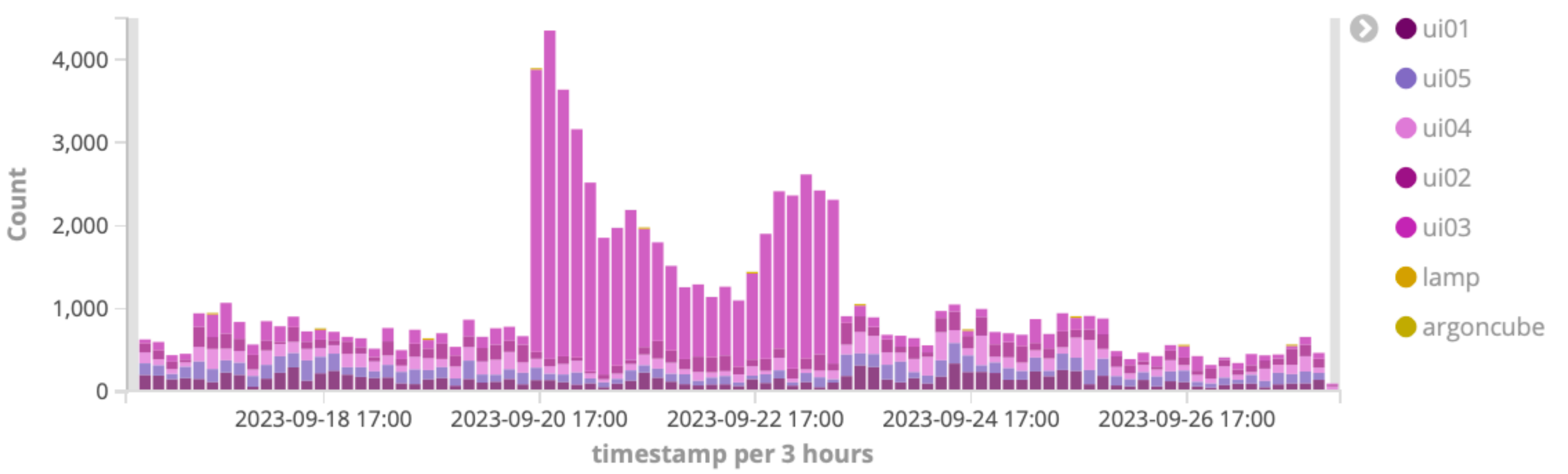
- **auditbeat** can help with system integrity and intrusion detection: we are exploring it
- we monitor **syslog** messages to spot anomalies, e.g. failed ssh attempts

sshd - Authentication Failure (stacked)

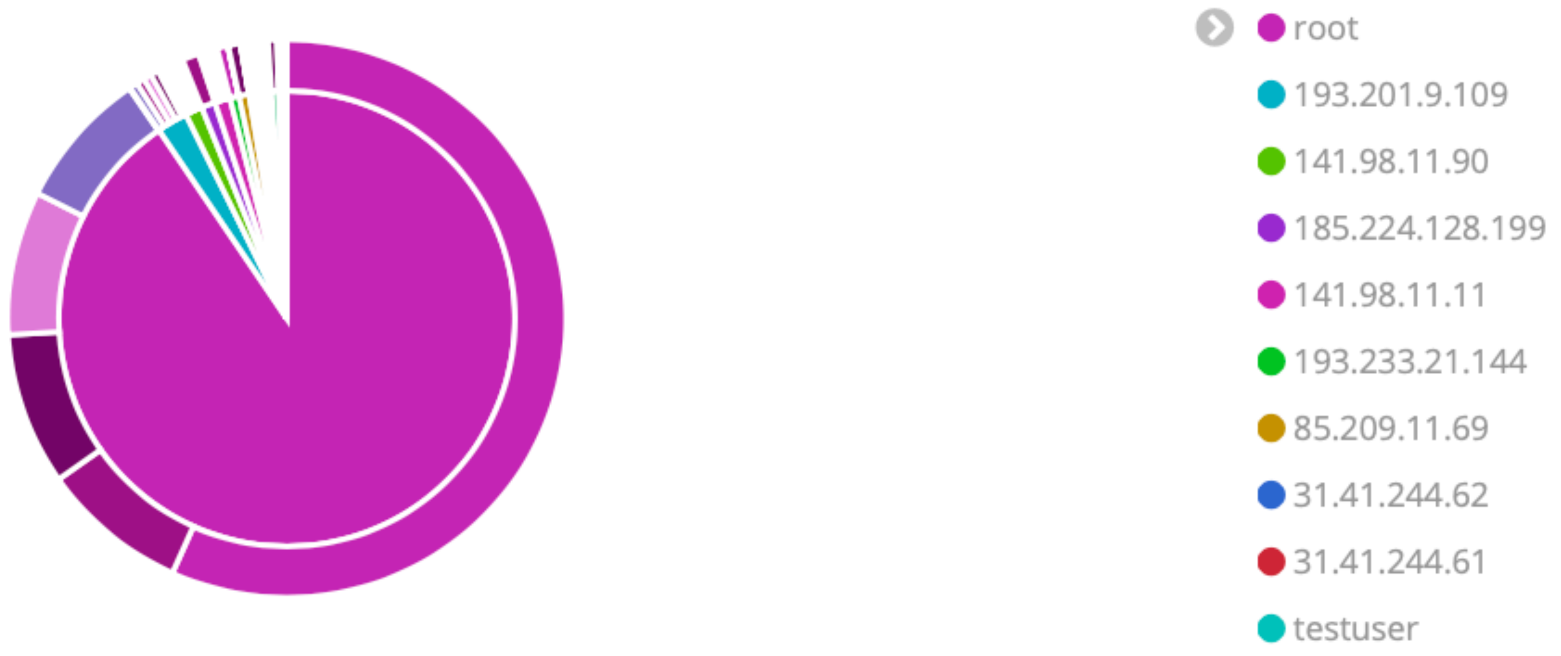


stuck fail2ban

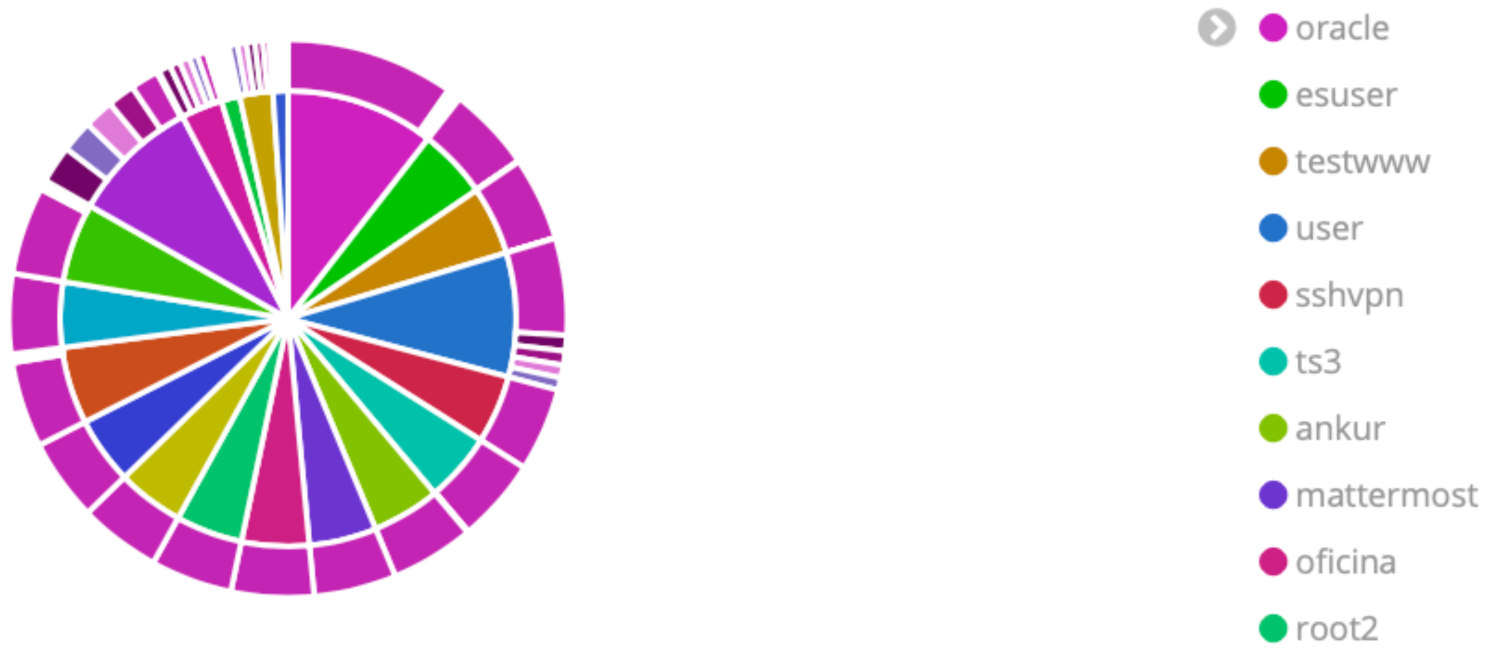
sshd - Invalid User Name (stacked)



sshd - Authentication Failure (pie)



sshd - Invalid User Name (pie)



OPERATING SYSTEMS & CONFIGURATION

- **Managed servers and cluster**
  - CentOS 7
    - cluster, CEs and lustre servers managed by Rocks
    - the rest is a mix of kickstart+postinstall and Ansible
  - Plan to transition to Alma 9 with Ansible
    - considering openHPC for the cluster and lustre
  - Ubuntu for the web server

ROCKS Open-Source Toolkit for Real and Virtual Clusters

- **Workstations, offices and labs (incl. University Hospital)**
  - A mix of CentOS, SL, Ubuntu, some Windows (DAQ, instrument control)
    - generally managed by the users, following first deployment
  - User laptops a mix of Mac OS, Ubuntu, Windows

PLANS

- **Cluster**
  - rolling replacement of older hardware ongoing
  - scale up to 15k slots
  - infiniband network re-factoring: dragonfly
- **Storage**
  - scale up lustre (*mds, flash pool*)
  - finalise accounting SRR for the federated dCache
  - migrate local DPM to dCache
- **OS**
  - Migrate to Alma 9 (*and re-benchmark with HS23*)
  - Rocks / OpenHPC / Ansible