

MIGRATING DPM TO THE FEDERATED NDGF-T1 DCACHE AT THE UNIBE-LHEP ATLAS TIER-2

Gianfranco Sciacca

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

HEPiX spring 2023 Taipei - 30 March 2023

u^b

^b
UNIVERSITÄT
BERN

AEC
ALBERT EINSTEIN CENTER
FOR FUNDAMENTAL PHYSICS

LABORATORIUM FÜR HOCHENERGIEPHYSIK
LHEP
UNIVERSITÄT BERN

WHO AND WHERE

▶ ATLAS Resource Centre @UniBE

- **LHEP** dedicated resource:
~10k cores, 0.6 PB **lustre** cache+scratch, 2.4 PB grid storage (0.5 PB for neutrinos, also CPUs)
- **UBELIX** @UniBE (multi-disciplinar cluster):
up to 2k cores opportunistically, **GPFS** cache+scratch
- **Baobab** @UniGE (multi-disciplinar cluster):
up to 500 cores opportunistically, **BeeGFS** scratch
- Up to 180 kHS06 (45 kHS06 opportunistic)

▶ Network

- 3 ms RTT from CERN (peering with Géant) and CSCS, 100Gbps backbone full redundant, 40Gbps sciDMZ



▶ **Three options for DPM migration - GDB Feb 2022**

- A. In place migration to dCache
- B. Consolidation of resources at the national level (dCache)
- C. Consolidation of resources at the international level (dCache)

▶ **Three options for DPM migration - GDB Feb 2022**

- A. In place migration to dCache
- B. Consolidation of resources at the national level (dCache)
- C. Consolidation of resources at the international level (dCache)

▶ **Decision taken at end of February 2022**

- C. International => Nordic T1

▶ **Three options for DPM migration - GDB Feb 2022**

- A. In place migration to dCache
- B. Consolidation of resources at the national level (dCache)
- C. Consolidation of resources at the international level (dCache)

▶ **Decision taken at end of February 2022**

- C. International => Nordic T1

▶ **Integration completed - GDB Sep 2022**

- ➔ PoC in place in January 2022
- ➔ Full migration completed by September 2022

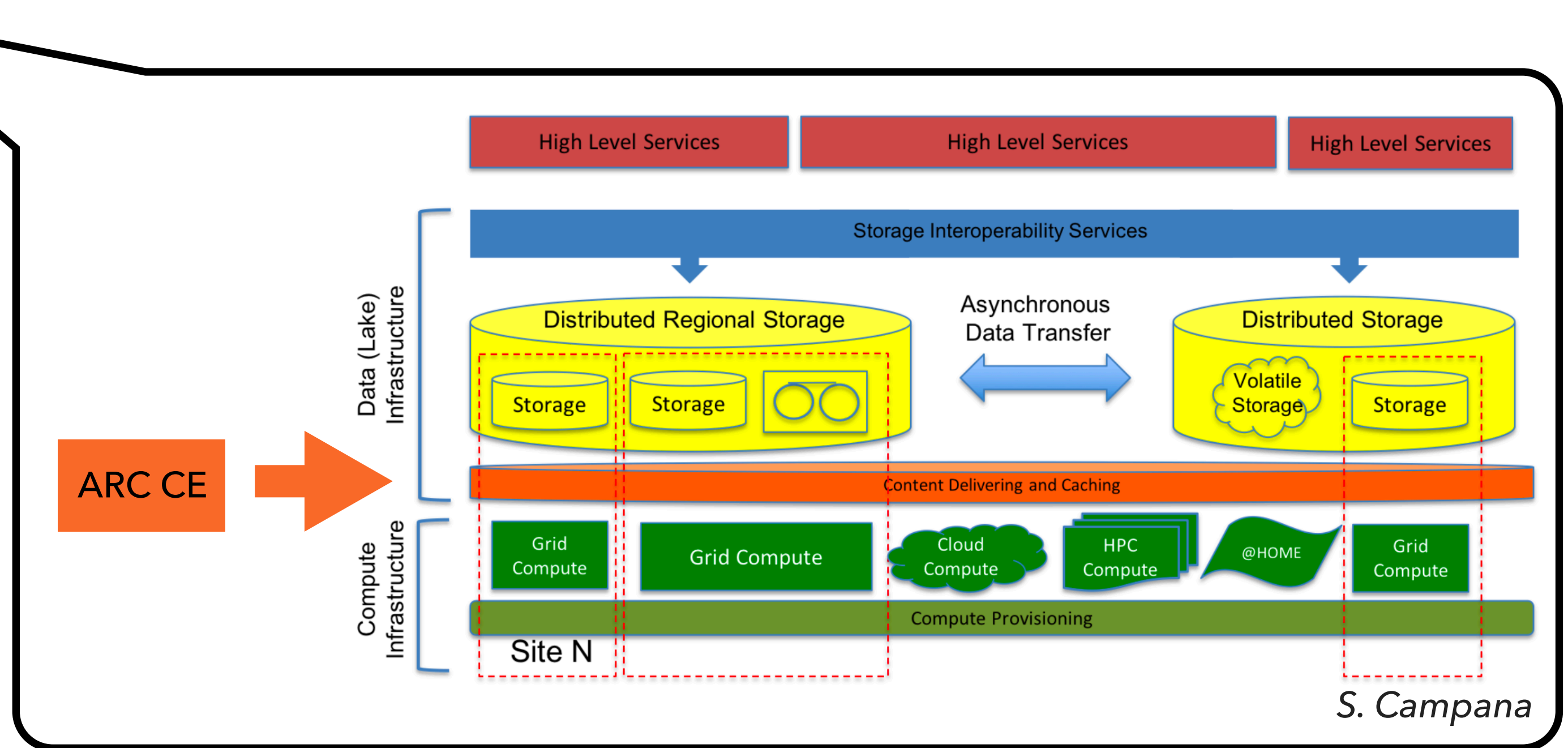
▶ Add value to the storage

- * Small storage: 1800TB pledged (ATLASDATADISK), ~100TB non-pledged (ATLASLOCALGROUPDISK)
- * Aim at providing higher value to researchers
- * In line with the WLCG data lake concept: **delocalised storage and low-latency local caches at computational centres**

MOTIVATION

► Add value to the storage

- * Small storage: 1800TB pledged (ATLASDATADISK), ~100TB non-pledged (ATLASLOCALGROUPDISK)
- * Aim at providing higher value to researchers
- * In line with the WLCG data lake concept: **delocalised storage and low-latency local caches at computational centres**



MOTIVATION

▶ Add value to the storage

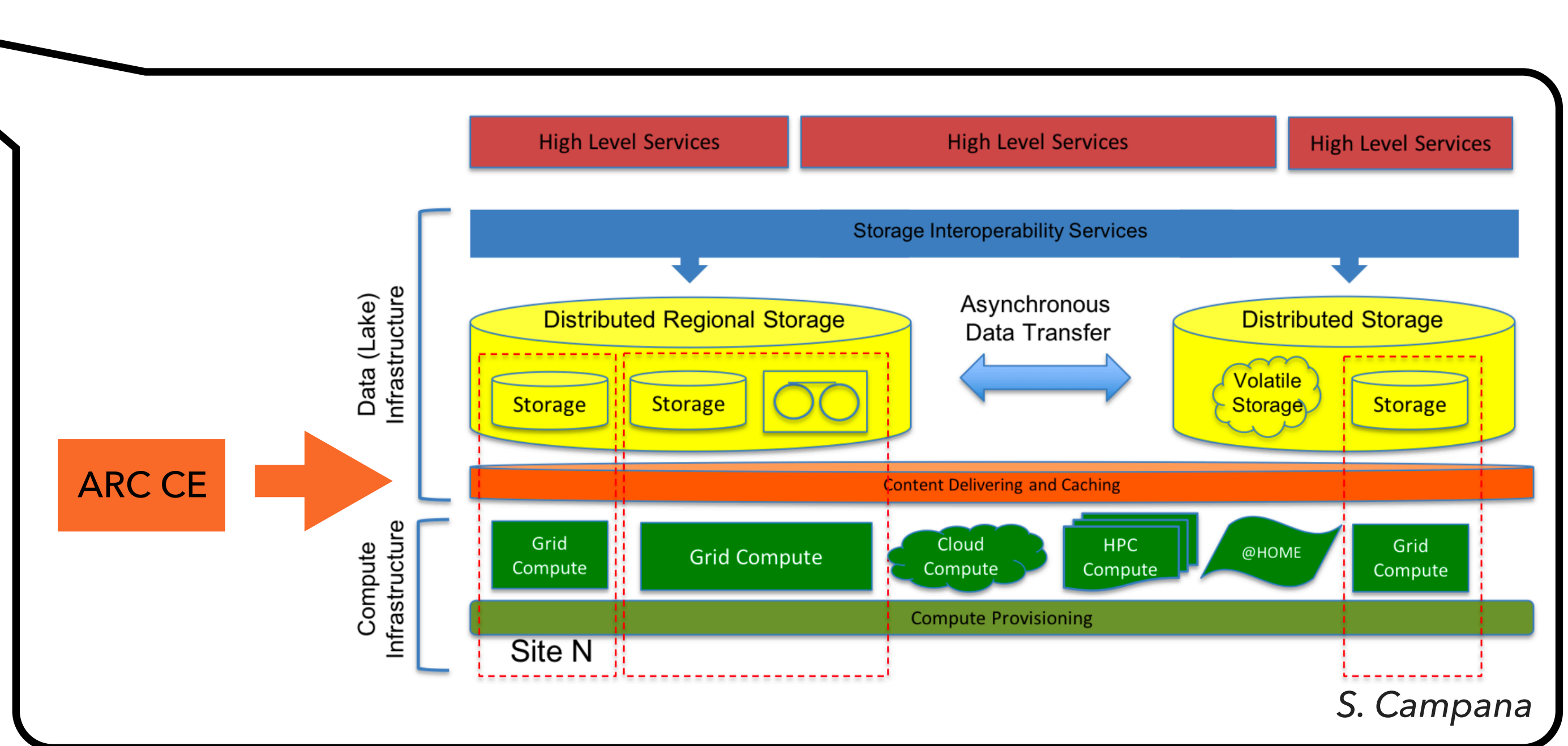
- * Small storage: 1800TB pledged (ATLASDATADISK), ~100TB non-pledged (ATLASLOCALGROUPDISK)
- * Aim at providing higher value to researchers
- * In line with the WLCG data lake concept: **delocalised storage and low-latency local caches at computational centres**

▶ Easy migration path

- * Well structured support

▶ Caveats

- * Increased pressure on the network
- * Potentially higher operational pressure



REMOTE SITE CHECKLIST

► Provision the storage

- * ATLAS data on the DPM drained by ATLAS central DDM team (*couple of weeks*)
- * Re-factorise data areas (*all servers were shared among other users*)
- * Hardware upgrades: NDGF checklist on recommended hardware configuration
- * Network upgrades => *interaction with the Uni NOC team*
 - * *10G or 2x10G on each server*
- * Fresh installation of servers with minimal CentOS 7
- * Tweaks to allow remote installation and management of dCache

REMOTE SITE CHECKLIST

▶ **Tuning, network and monitoring** (*wiki checklists*)

- * Apply OS and TCP tuning as recommended for performance (*sysctl*)
 - * *TCP window for high RTT transfers*
 - * *BBR congestion control, vm.swappiness, vm.min_free_kbytes, ...*
- * Firewall / ACLs settings
 - * *Allow ssh, ganglia, prometheus, dCache data*
- * Integration with centralised monitoring @NDGF
 - * *Ganglia / Prometheus*
 - * *Local monitoring at the server level (inc. network) at the remote site*

▶ **Handover => NDGF admins**

NDGF ADMIN CHECKLIST

▶ dCache deployment and commissioning

- * dCache stack deployed via Ansible and an unprivileged user account
 - * *“tarpool” (tar file distribution as opposed to a deb/rpm package to install locally)*
 - * *Install and upgrade Java and dCache and apply the required configuration of dCache*
 - * *Unprivileged access via ssh keys*

- * Ensure communication with the remote pools can be established, add to the monitoring

NDGF ADMIN CHECKLIST

▶ dCache deployment and commissioning

- * Load-test pools before taking them to production
 - * *Fill them up with production data from nearby/fast pools, then checksum all data on disk*
 - * *See how fast you can read data out (migration cache to a different set of pools)*
 - * *Migrate in some new data, see if writes starves reads, or vice versa.*
 - * *Add to a read-only pool group to get real client reads from the pool*

- * **Add to production poolgroup**

▶ **Structured communication**

- * thematic chat rooms: general ops, middleware, site-support, etc ...
- * operations, support, incidents, ...
- * operator on duty/on call
- * co-ordinate downtimes, upgrades, central and at sites, generally outage-free
- * dCache upgrade performed centrally for all remote pools, remote site local work sync-ed to minimise disruption

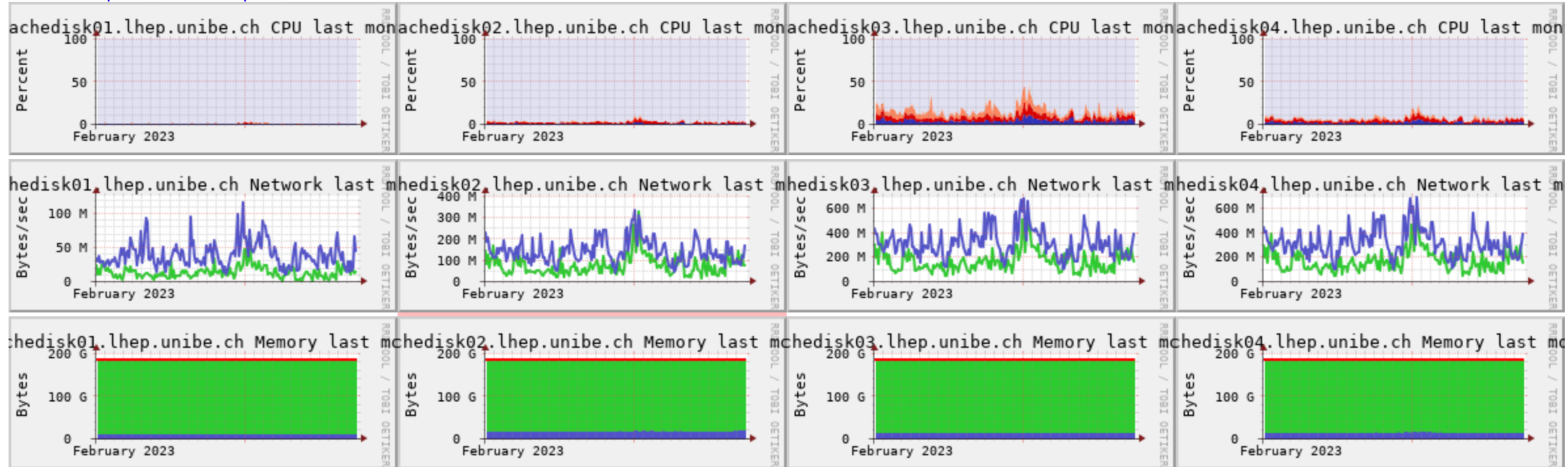
▶ **Regular meet-ups**

- * weekly ops online
- * topical meetings
- * face to face twice a year

IN PRODUCTION (CENTRAL VIEW)

UniBe pools













[Show UniBe pools](#) [Hide UniBe pools](#)



Service Overview For Host Group 'site-unibe-pools'

Host	Domain	IP	Port	Services	Health
lhep_unibe_ch_001	dcachedisk01_lhep_unibe_ch_01Domain	79691776	2451870	0	■ ■ ■
lhep_unibe_ch_002	dcachedisk02_lhep_unibe_ch_1Domain	94371840	3881298	0	■ ■ ■
lhep_unibe_ch_003	dcachedisk02_lhep_unibe_ch_2Domain	133169152	5153440	0	■ ■ ■
lhep_unibe_ch_004	dcachedisk02_lhep_unibe_ch_3Domain	94371840	4125608	0	■ ■ ■
lhep_unibe_ch_005	dcachedisk03_lhep_unibe_ch_1Domain	359661568	16542025	0	■ ■ ■
lhep_unibe_ch_006	dcachedisk03_lhep_unibe_ch_2Domain	359661568	16774711	0	■ ■ ■
lhep_unibe_ch_007	dcachedisk04_lhep_unibe_ch_1Domain	358612992	16890839	0	■ ■ ■
lhep_unibe_ch_008	dcachedisk04_lhep_unibe_ch_2Domain	358612992	17304732	0	■ ■ ■

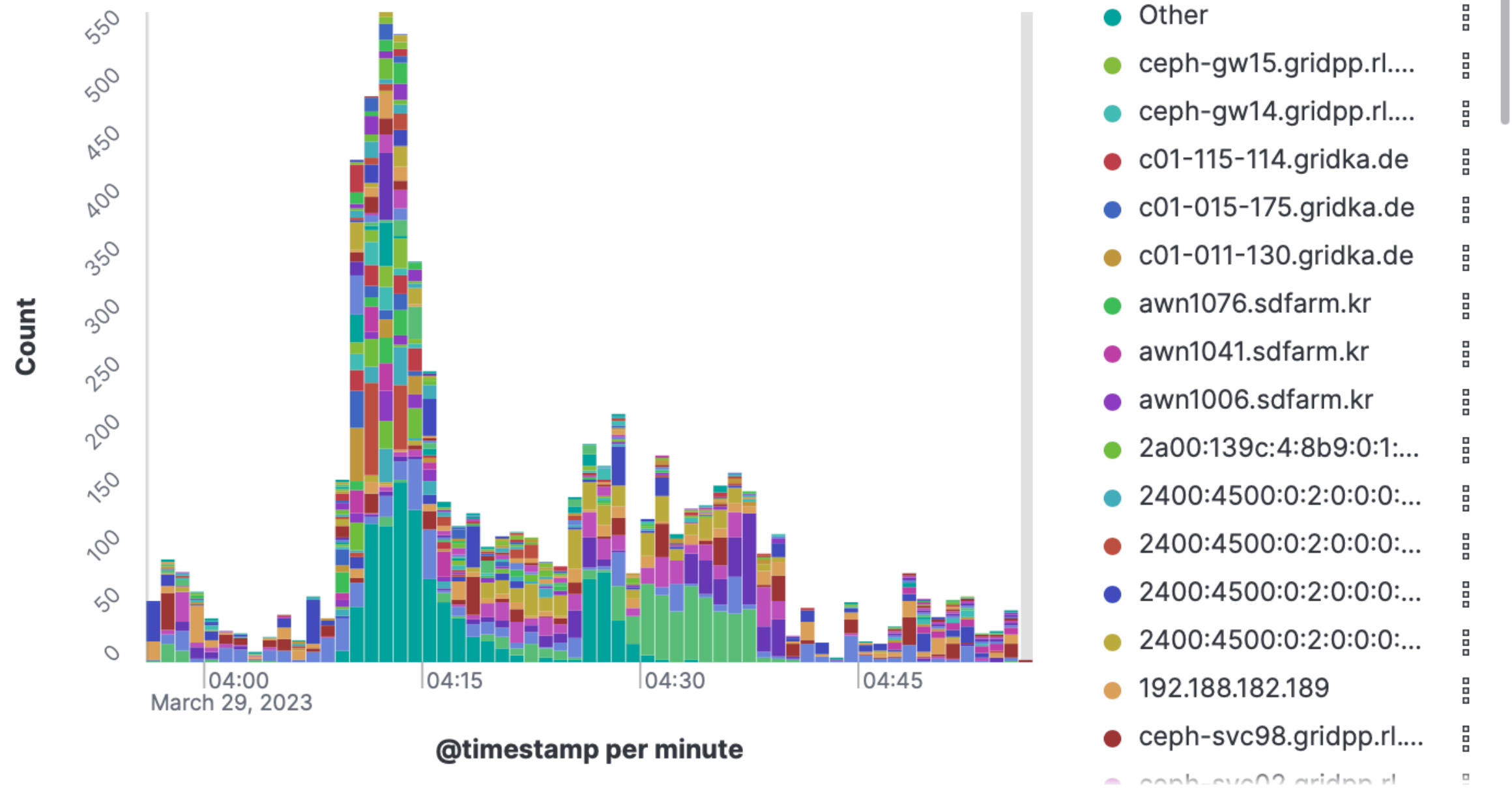
UNIBE dCache pools (site-unibe-pools)

Host	Status	Services	Actions
dcachedisk01.lhep.unibe.ch	UP	1 OK	  
dcachedisk02.lhep.unibe.ch	UP	1 OK	  
dcachedisk03.lhep.unibe.ch	UP	1 OK	  
dcachedisk04.lhep.unibe.ch	UP	1 OK	  

IN PRODUCTION (CENTRAL VIEW)

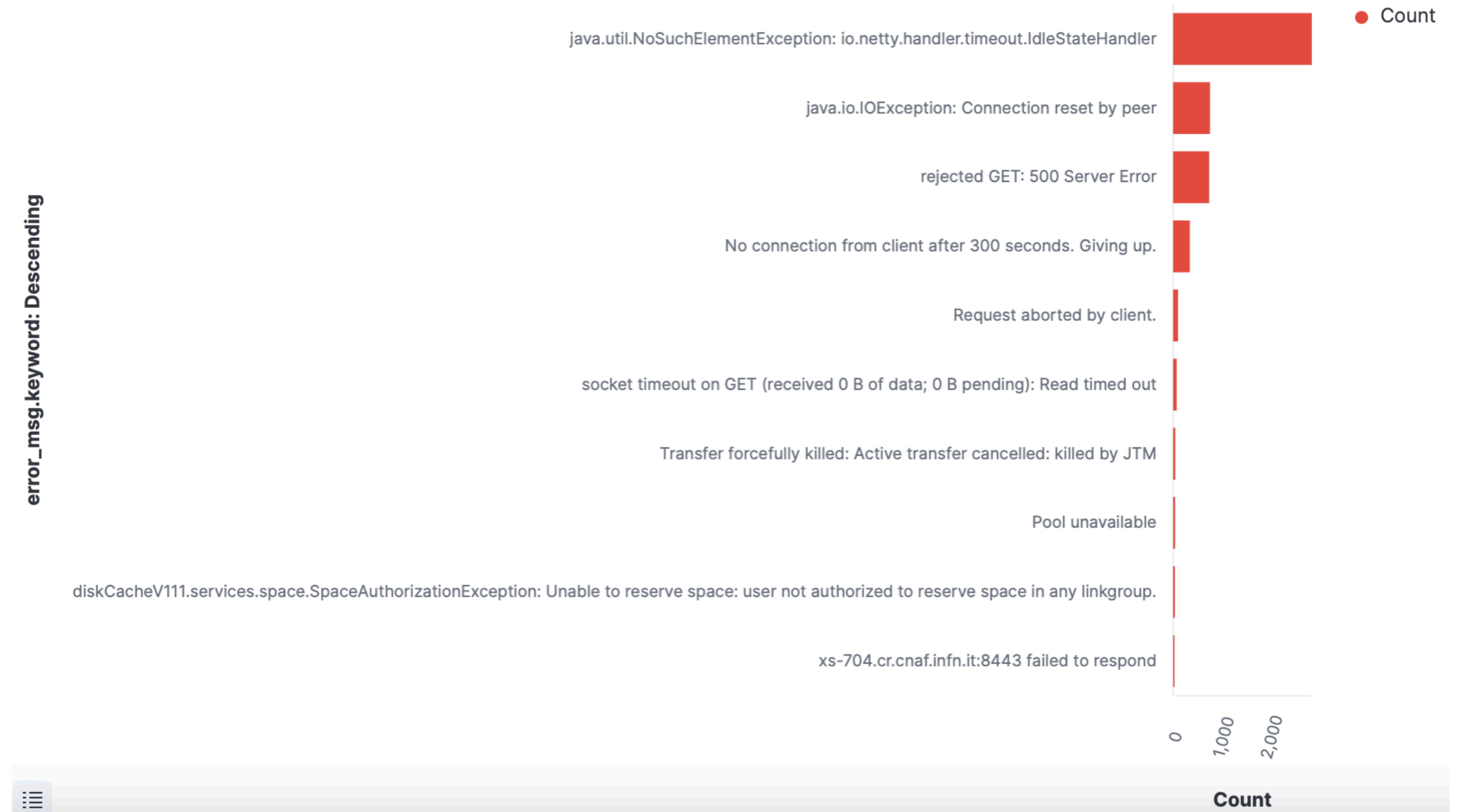
Count by `remote_hostname.keyword`

NOT error_code.keyword: 0 ×



Count by `err_msg.keyword`

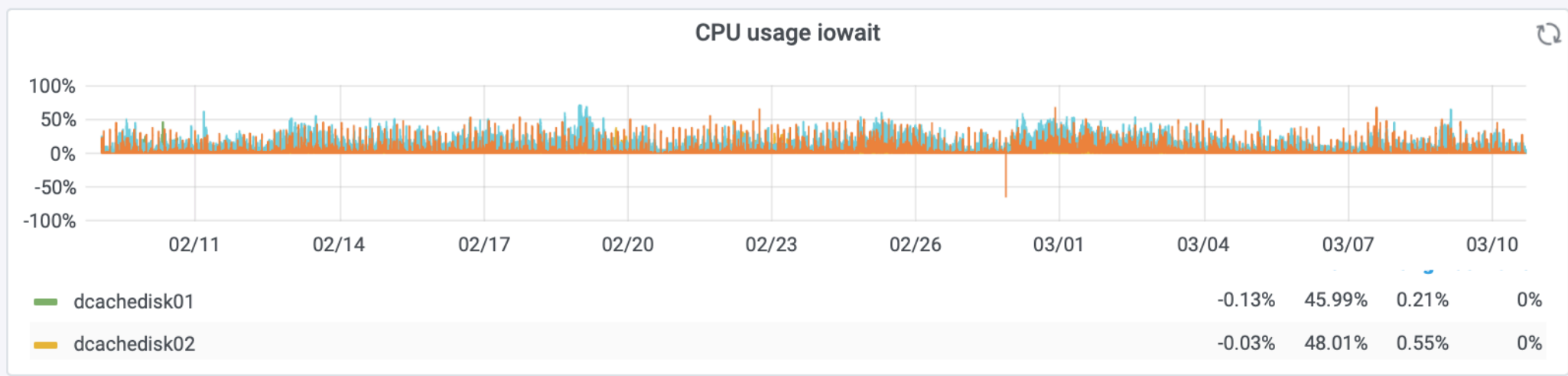
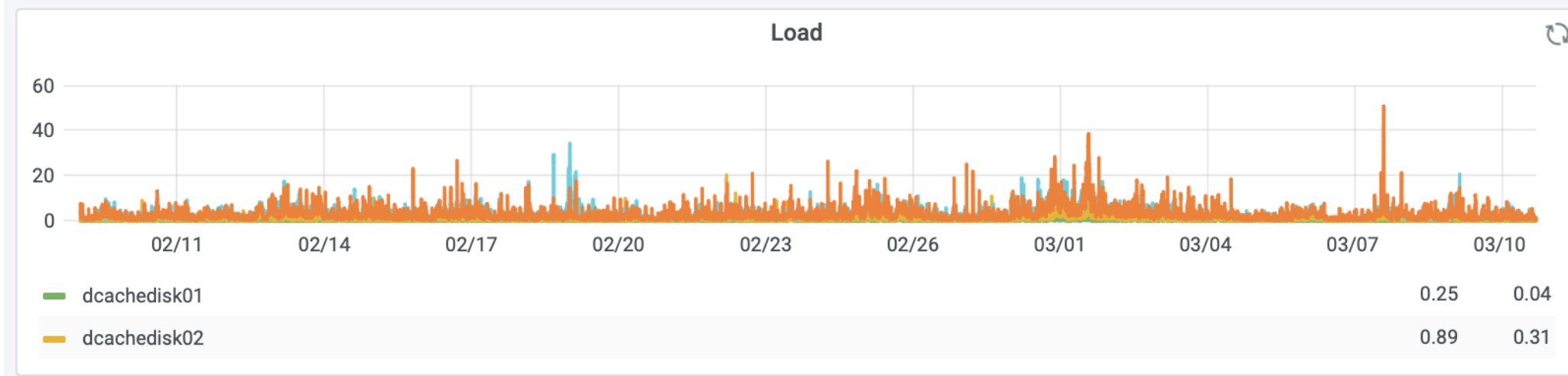
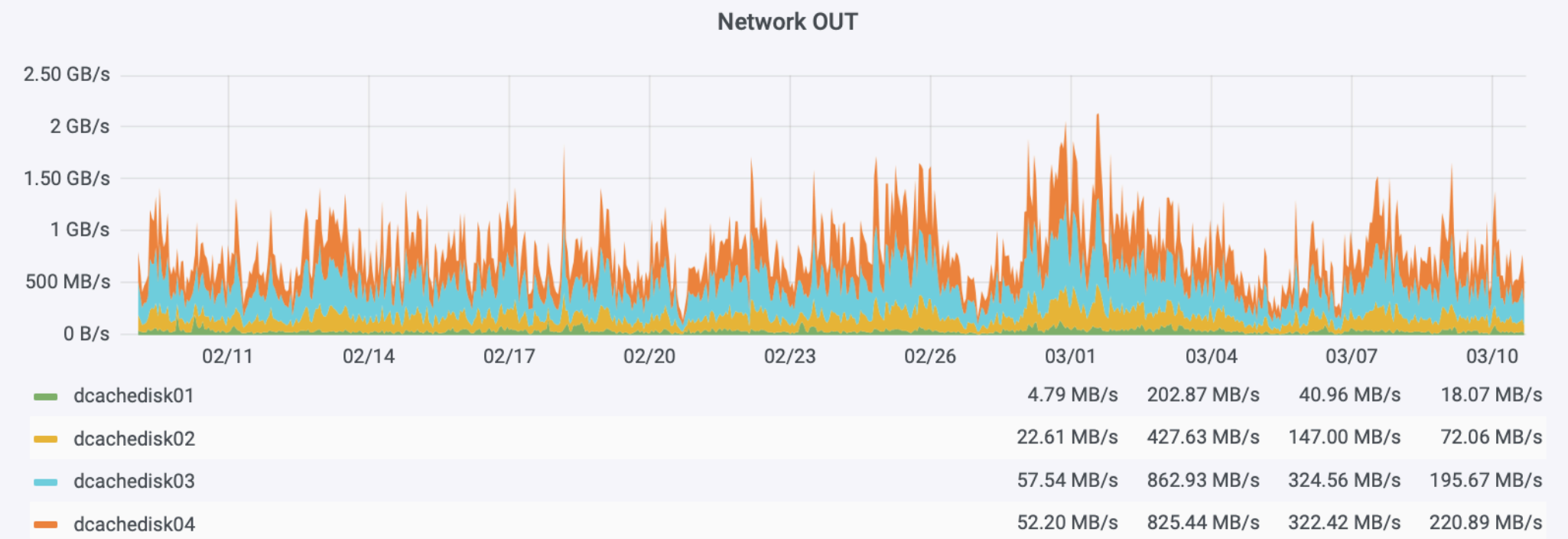
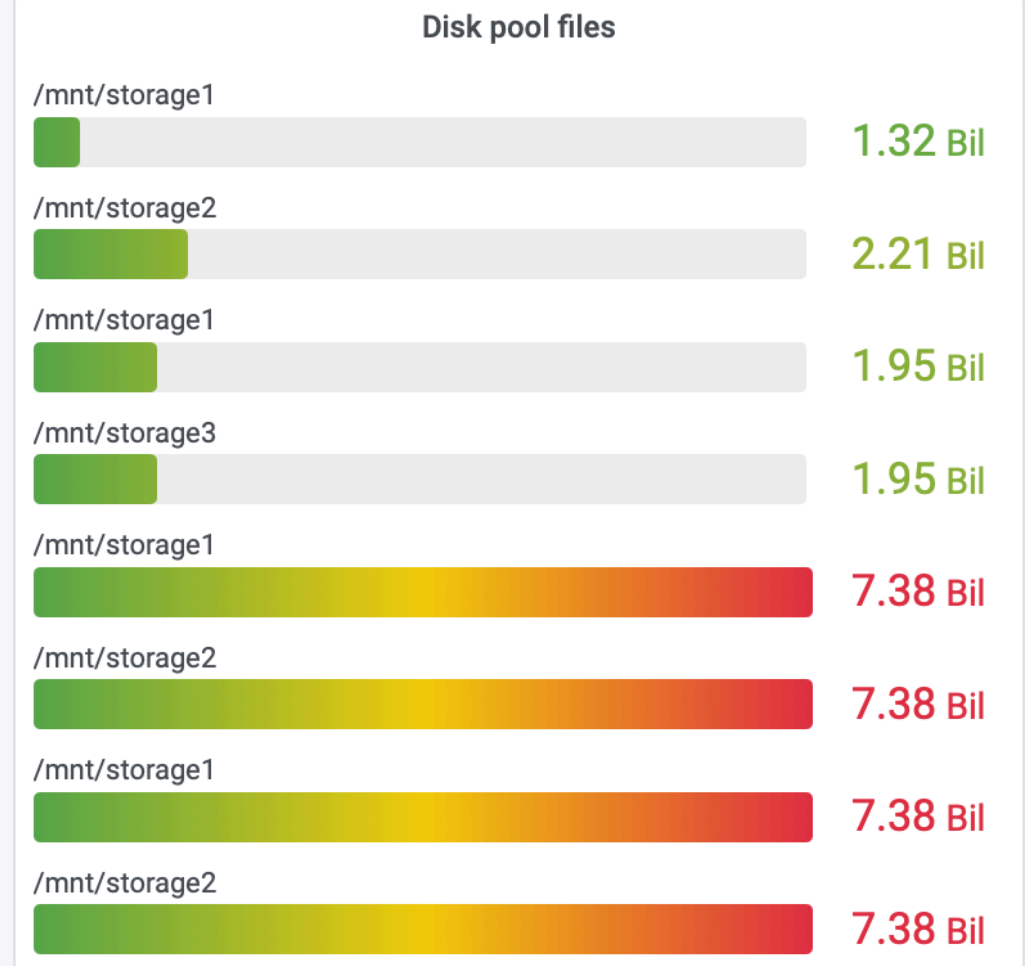
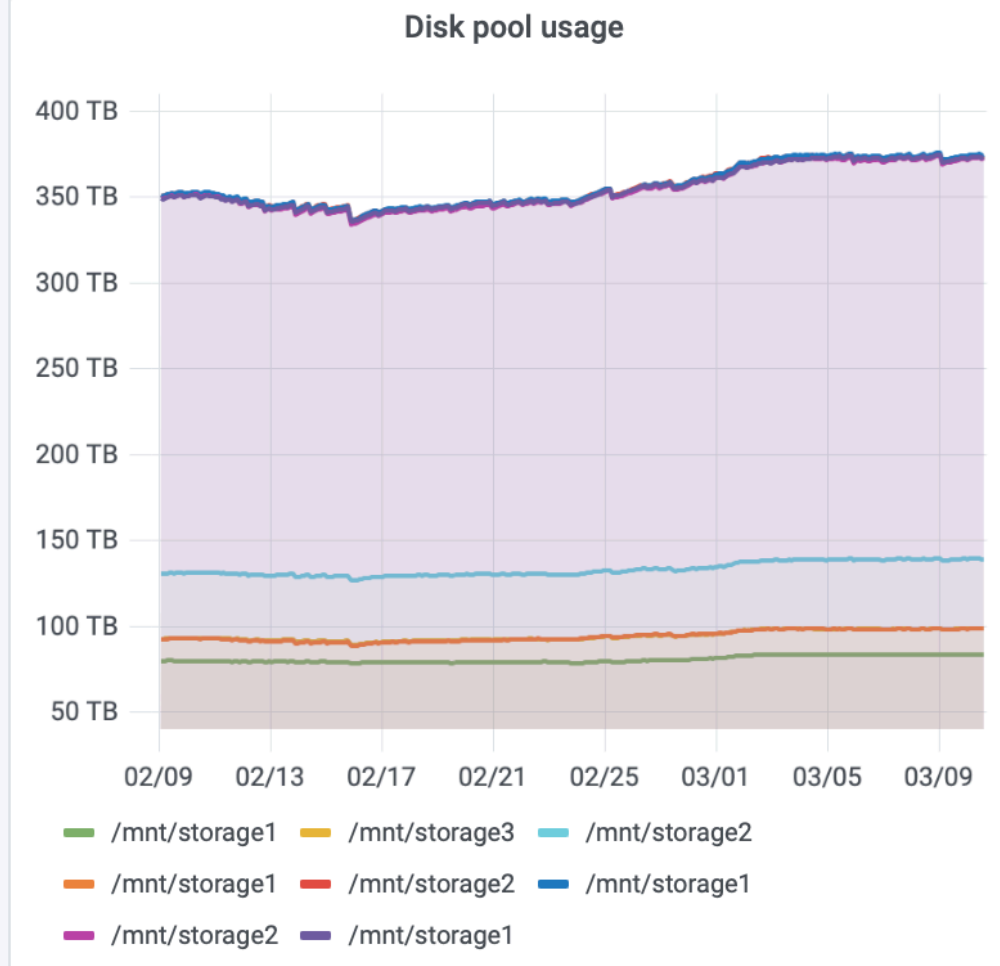
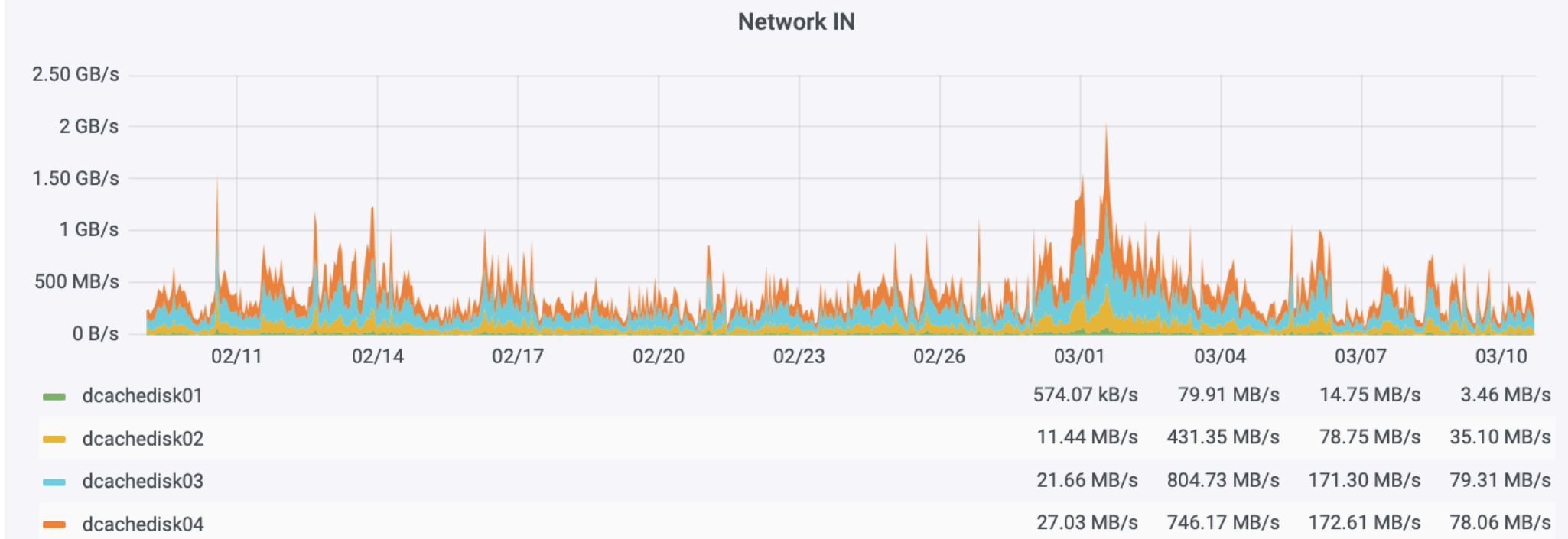
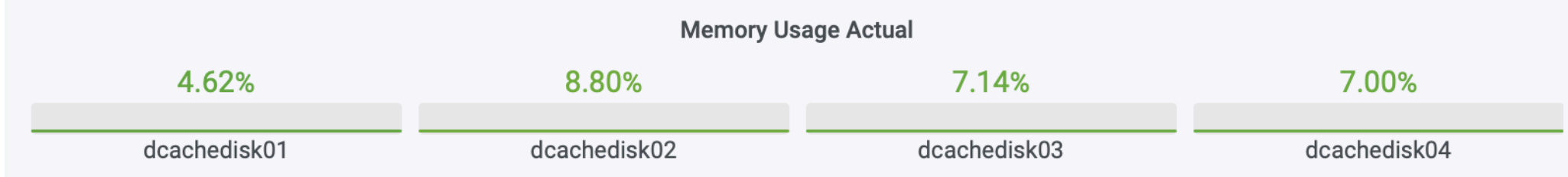
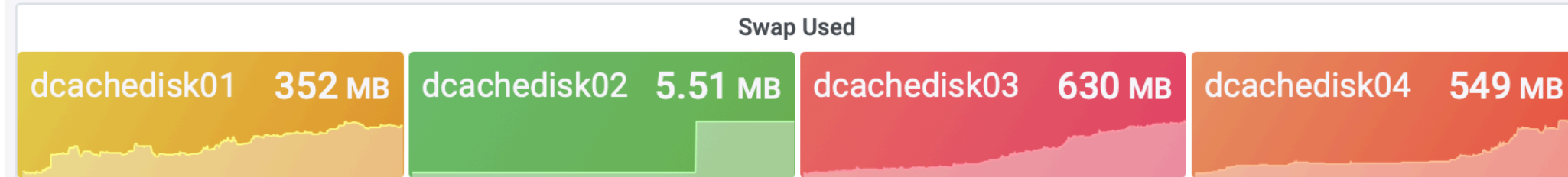
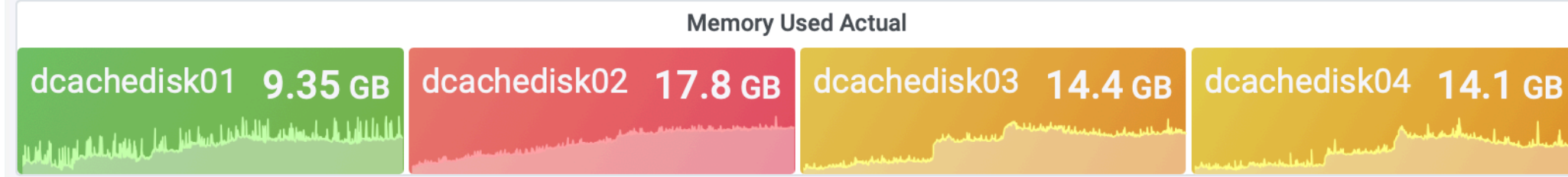
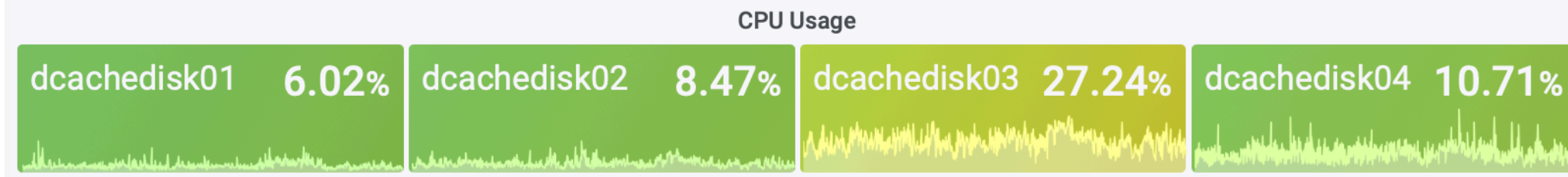
NOT error_code.keyword: 0 ×



IN PRODUCTION (SITE VIEW)

LHEP Services / dCache pools ☆ 🔗

📊 📄 ⚙️ 🕒 Last 30 days 🔍 🔄 5m 🗨️



▶ Storage Resource Reporting (SRR) for federated storage

- * must allow tracking of remote site storage contribution in the monthly WLCG/CRIC reports
 - * discussed and agreed on at the WLCG ops coordination meeting in July 2022
 - * additional share per site to the SSR, e.g. : `xxx_admin_UNIBE-LHEP`
 - * the WSSA (WLCG storage space accounting) application should handle these shares accordingly to account the corresponding space to the correct site
 - * the experiment operation is instead only concerned about the whole NDGF storage area

- * In progress now

CONCLUSIONS

- ▶ **The ATLAS T2 DPM storage @UNIBE-LHEP has been integrated with the Nordic T1 dCache**
 - * third Tier-2 site storage integrated with NDGF
 - * streamlined integration procedure
 - * smooth operation in production
 - * accounting of remote storage for WLCG is being finalised