

# Operating a distributed dCache

NeIC NT1 Manager  
Mattias Wadenstein  
<[maswan@ndgf.org](mailto:maswan@ndgf.org)>

2023-03-30  
HEPiX  
Taipei, Taiwan



# Overview

- Organization
- Service description
- Monitoring
- Ansible
- dCache admin



# Organization

## NeIC NT1 central operations

- Staff on NeIC contracts
- Distributed team at partner institutes
- Lead by NT1 manager

## • Site operations

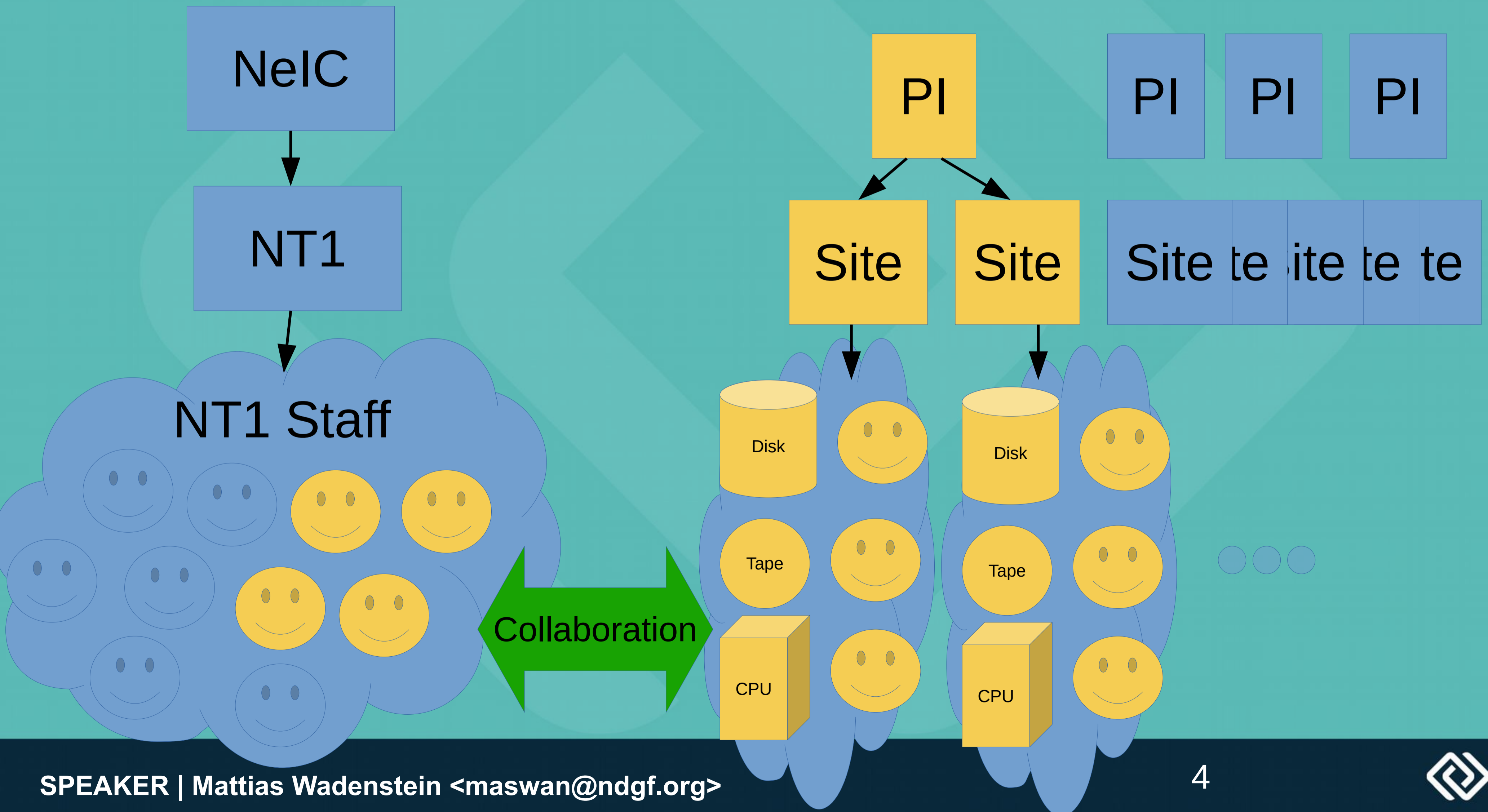
- 6 Nordic sites and 2 extended
- Owns and runs the hardware
- Provides managed services (batch systems, tape archives, etc)

## • Plus external services, like networking



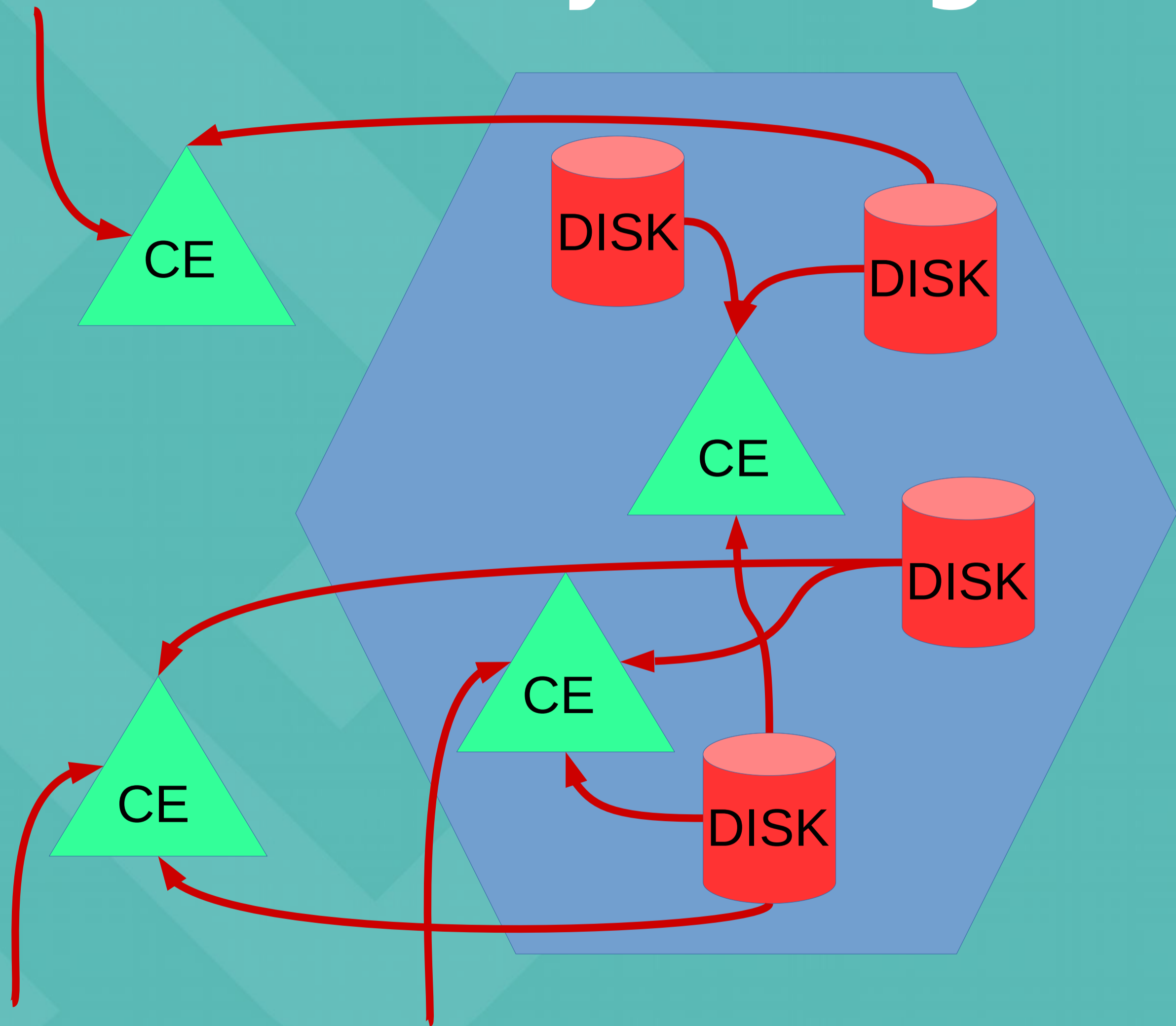


# Organization

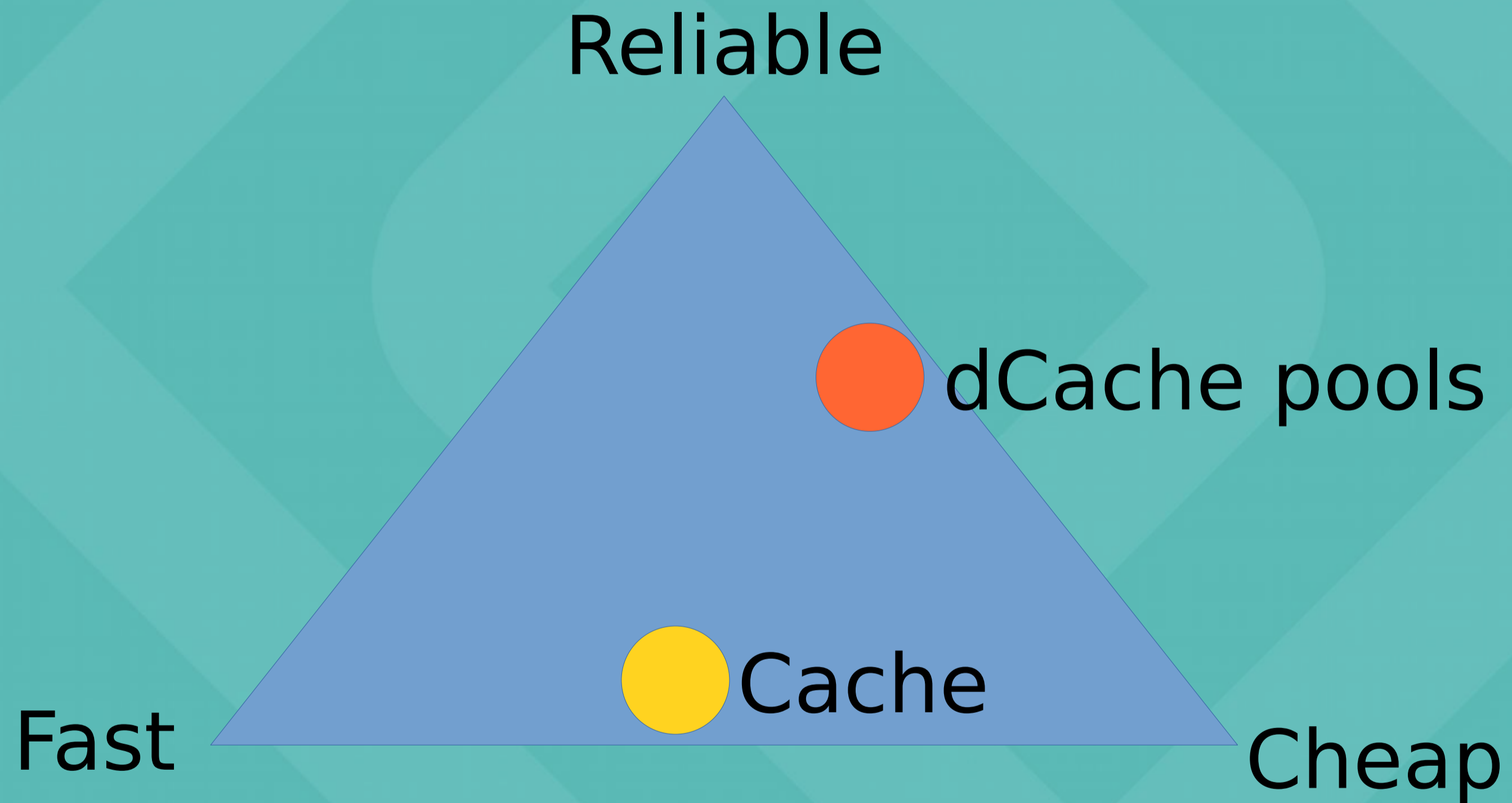


# ARC caching for latency hiding

- ARC is location agnostic
- No problem getting some data from other sites
- By staging data before running jobs, good CPU efficiency

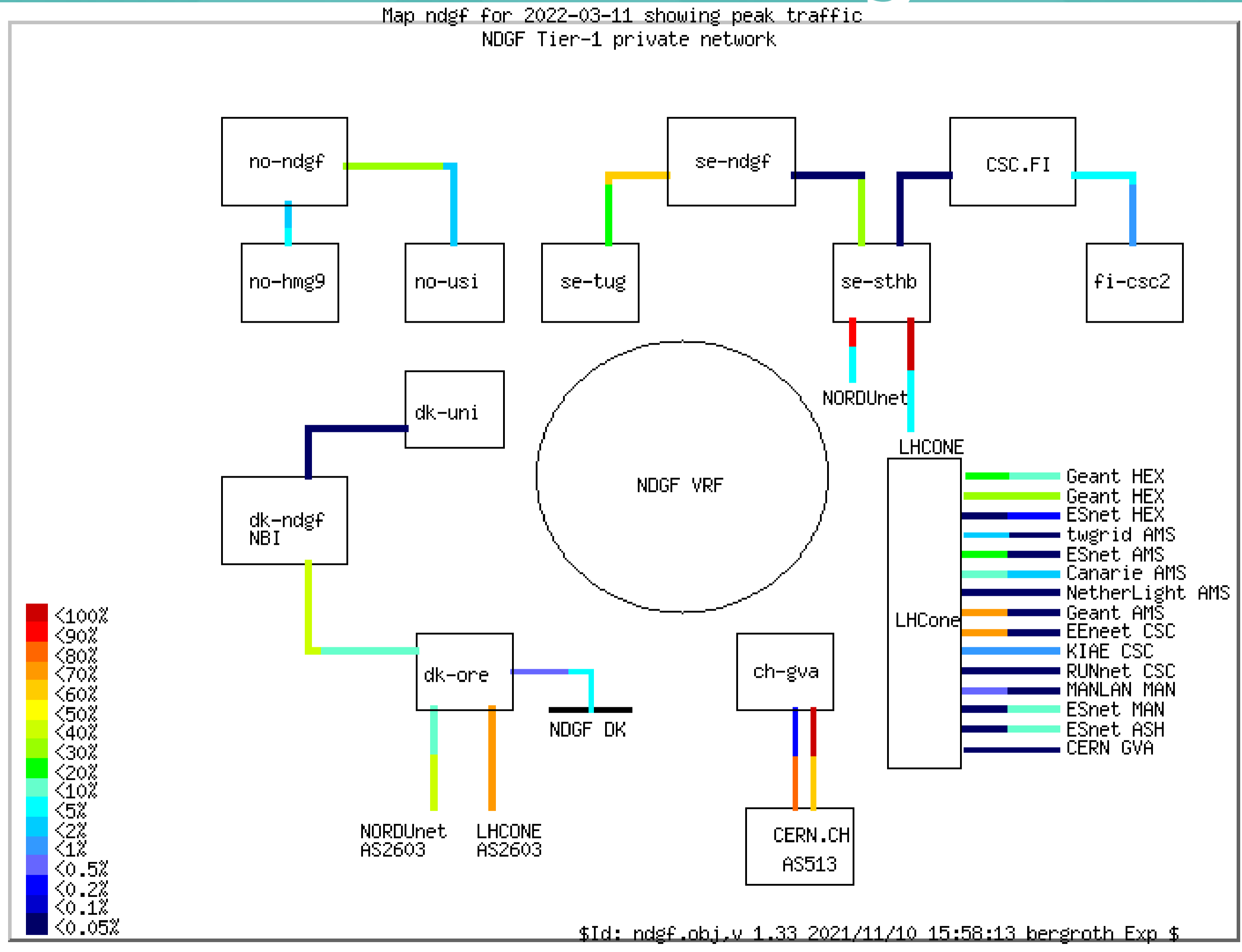


# ARC for Caching

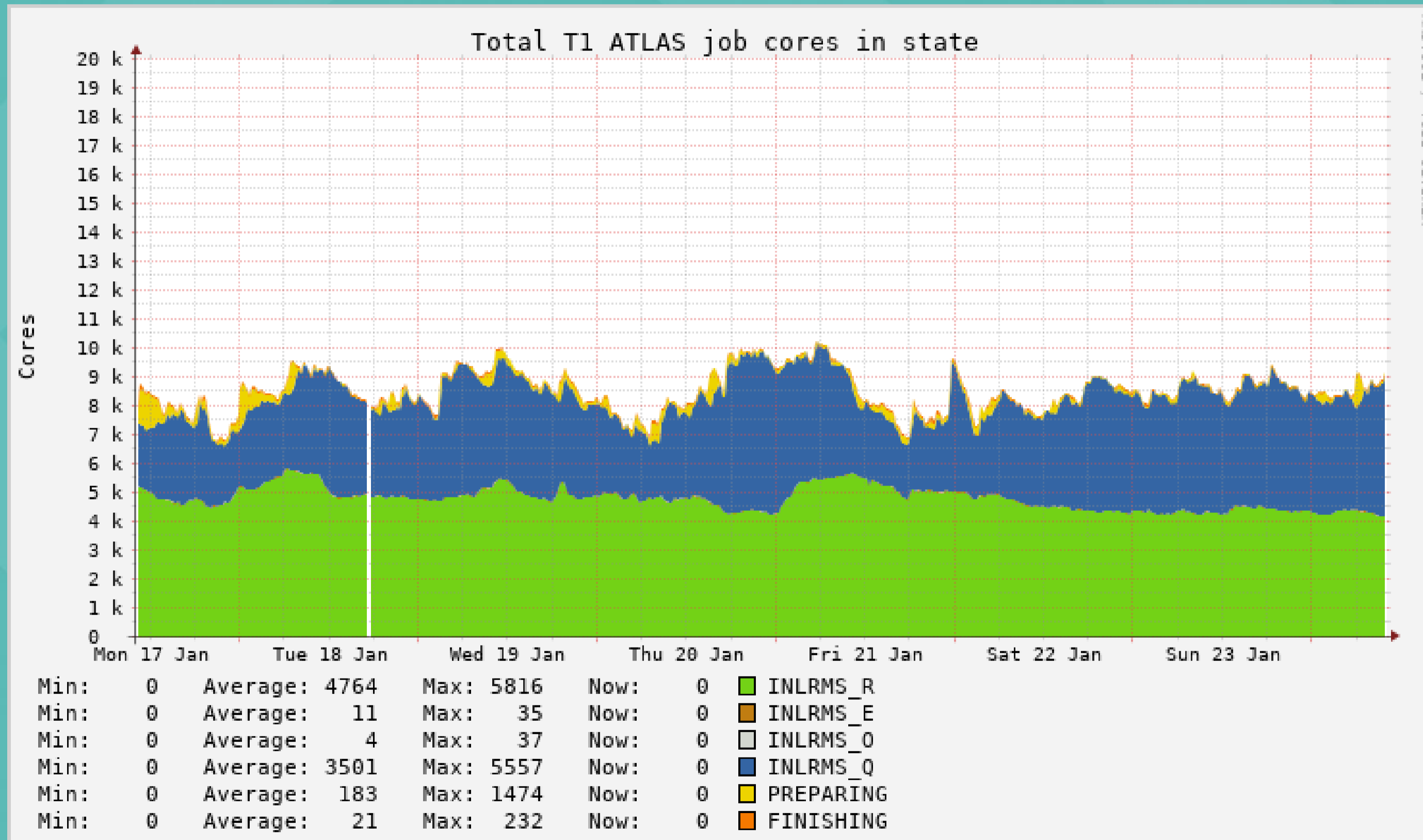


# Networking

Map ndgf for 2022-03-11 showing peak traffic  
 NDGF Tier-1 private network

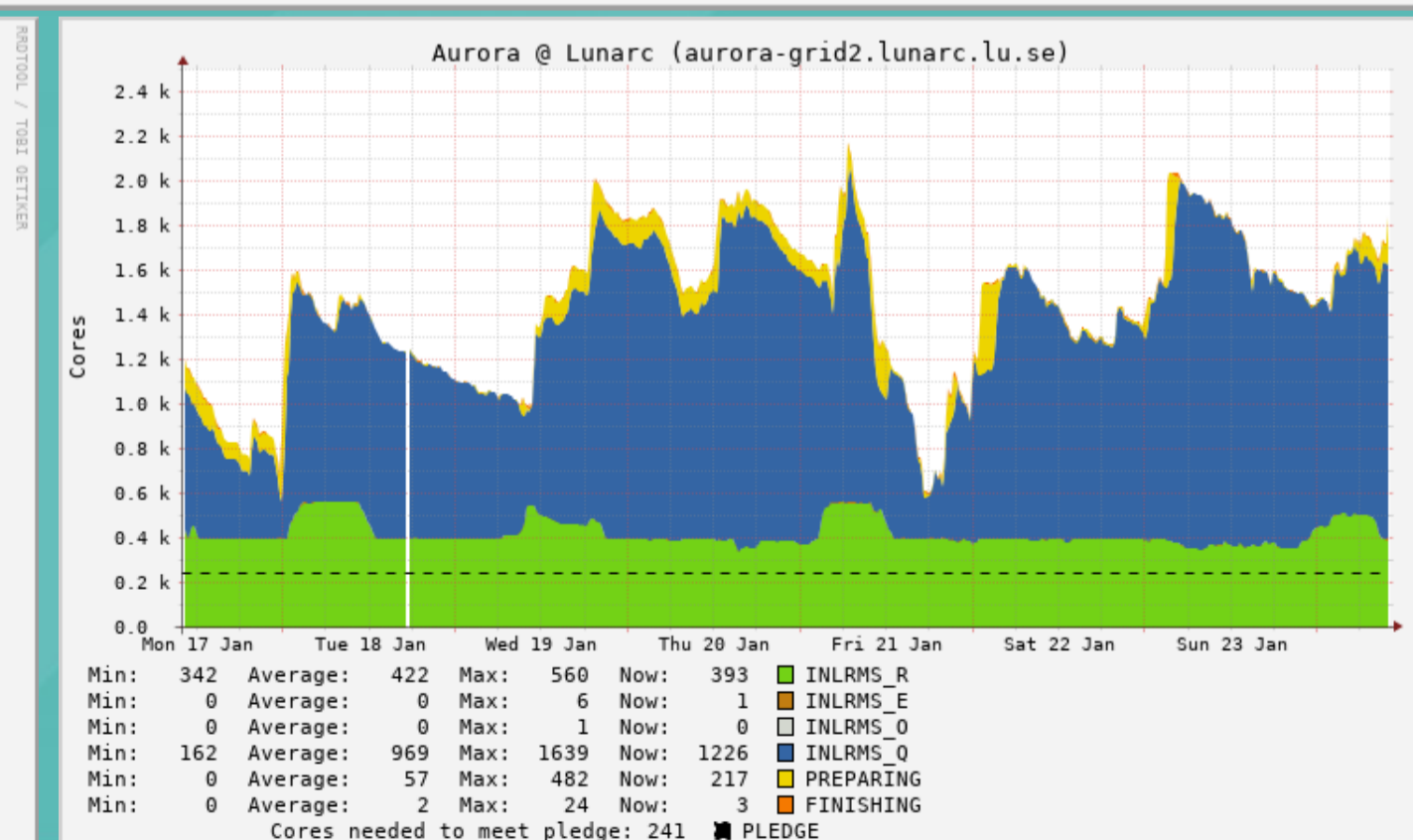
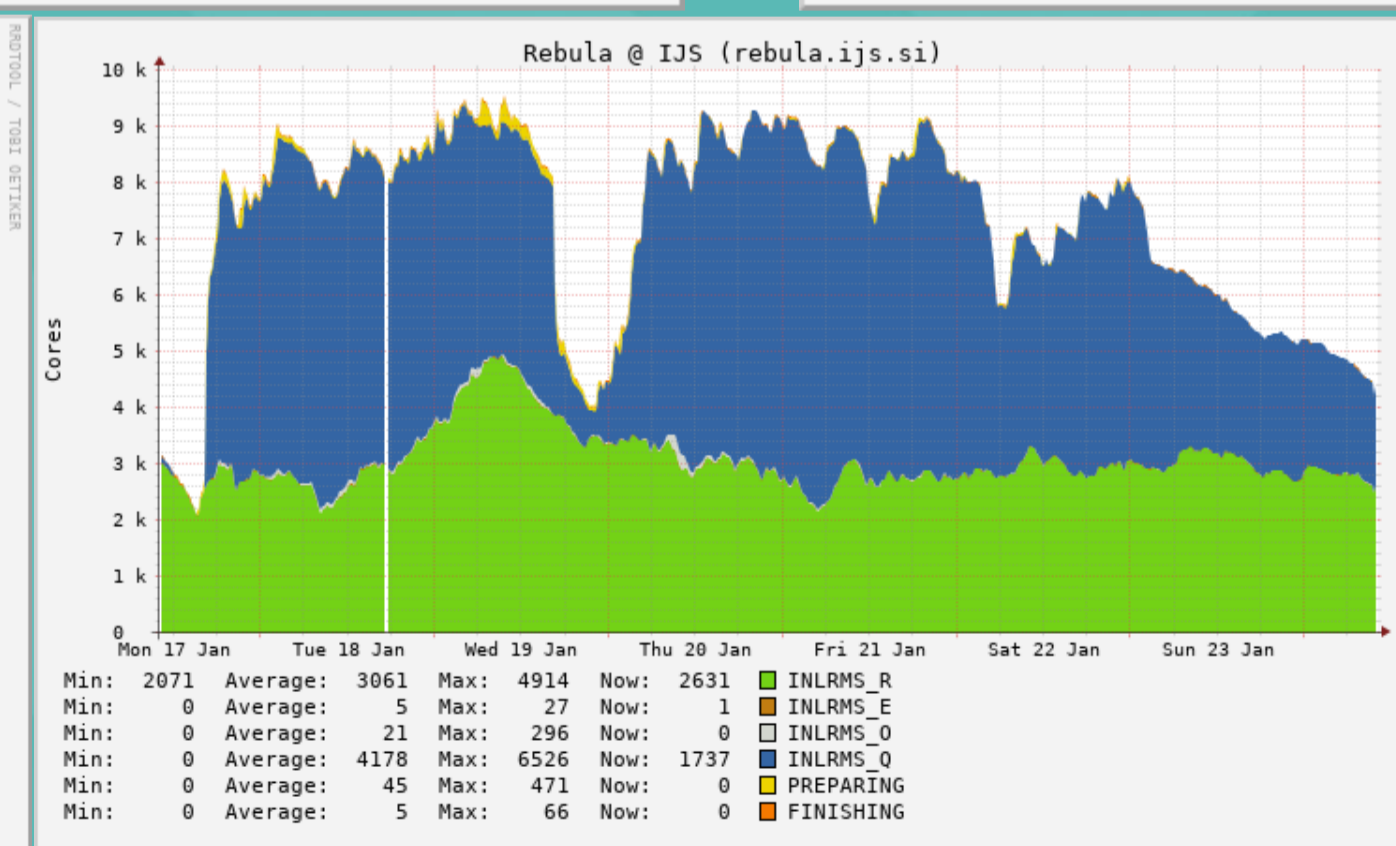
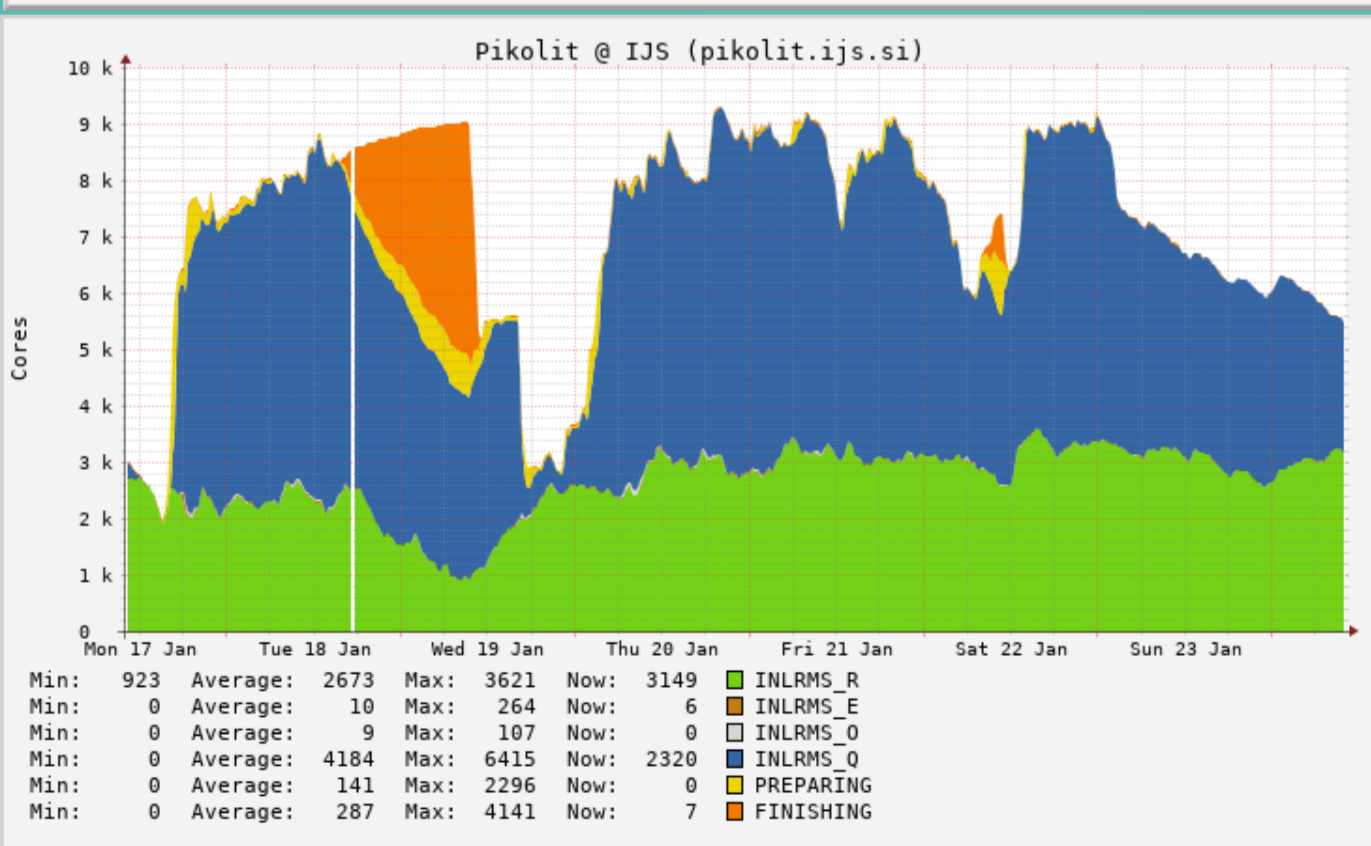
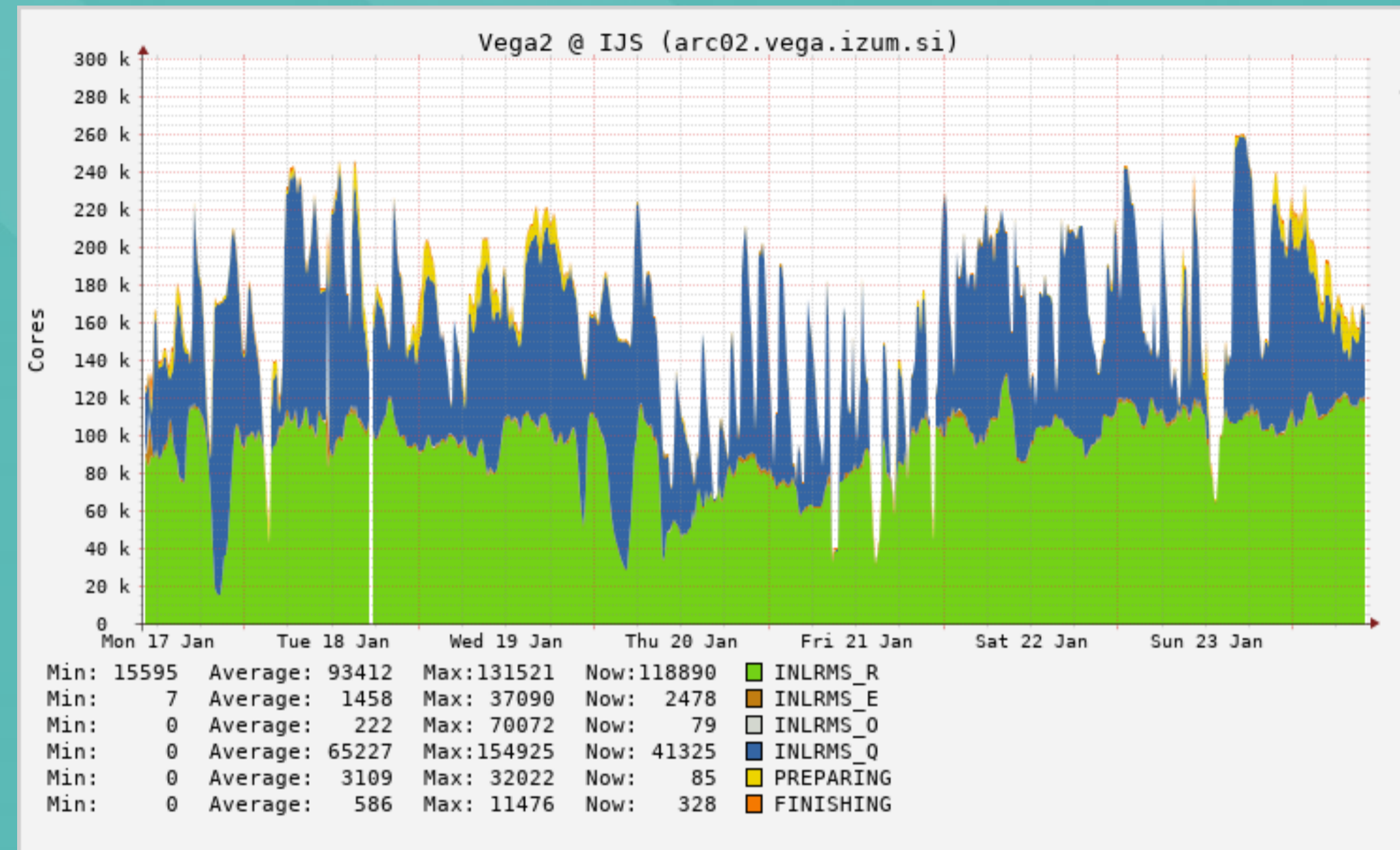
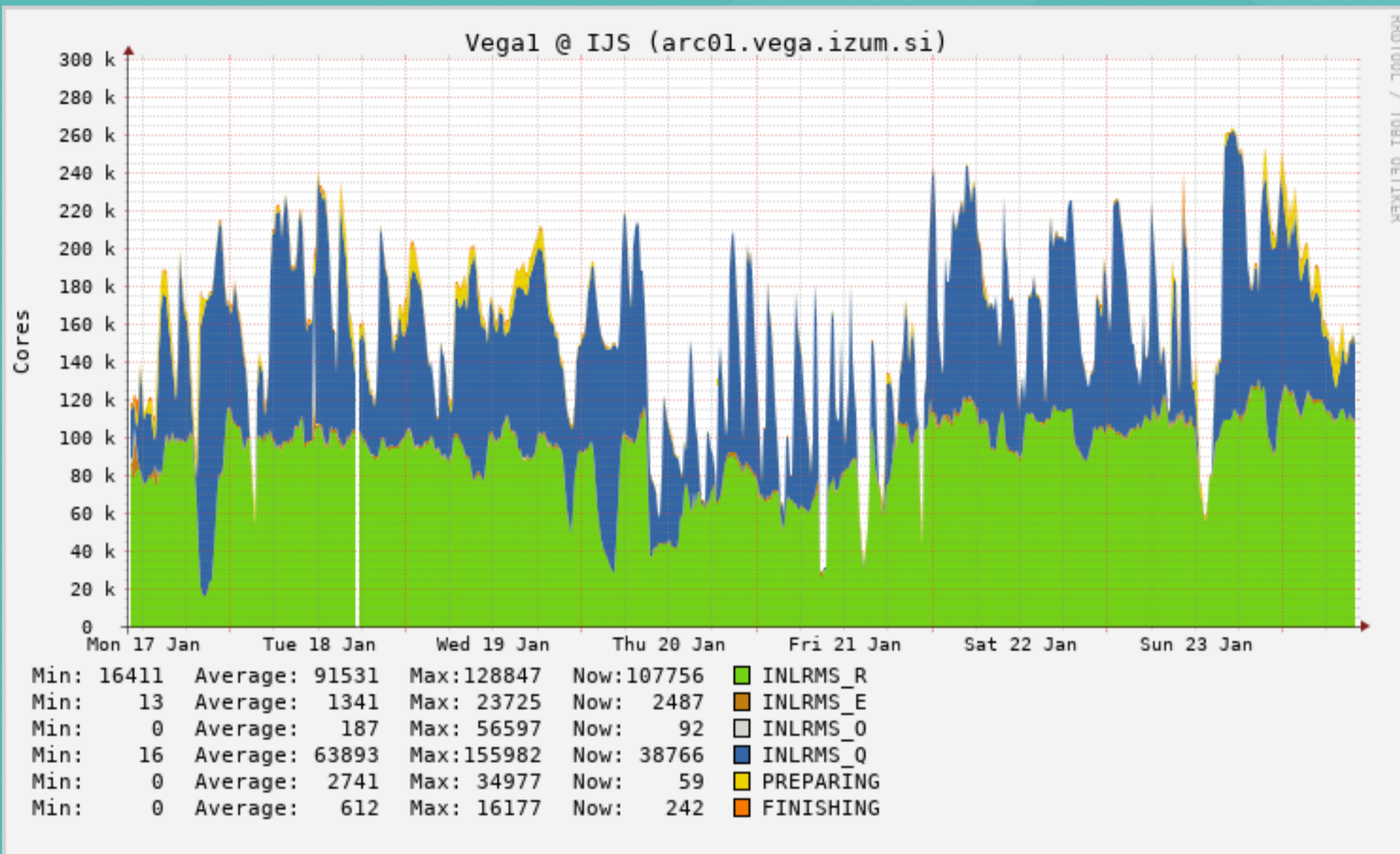


# CPU, tier1-share only





# CPU, associated tier-2s





# Storage service overview

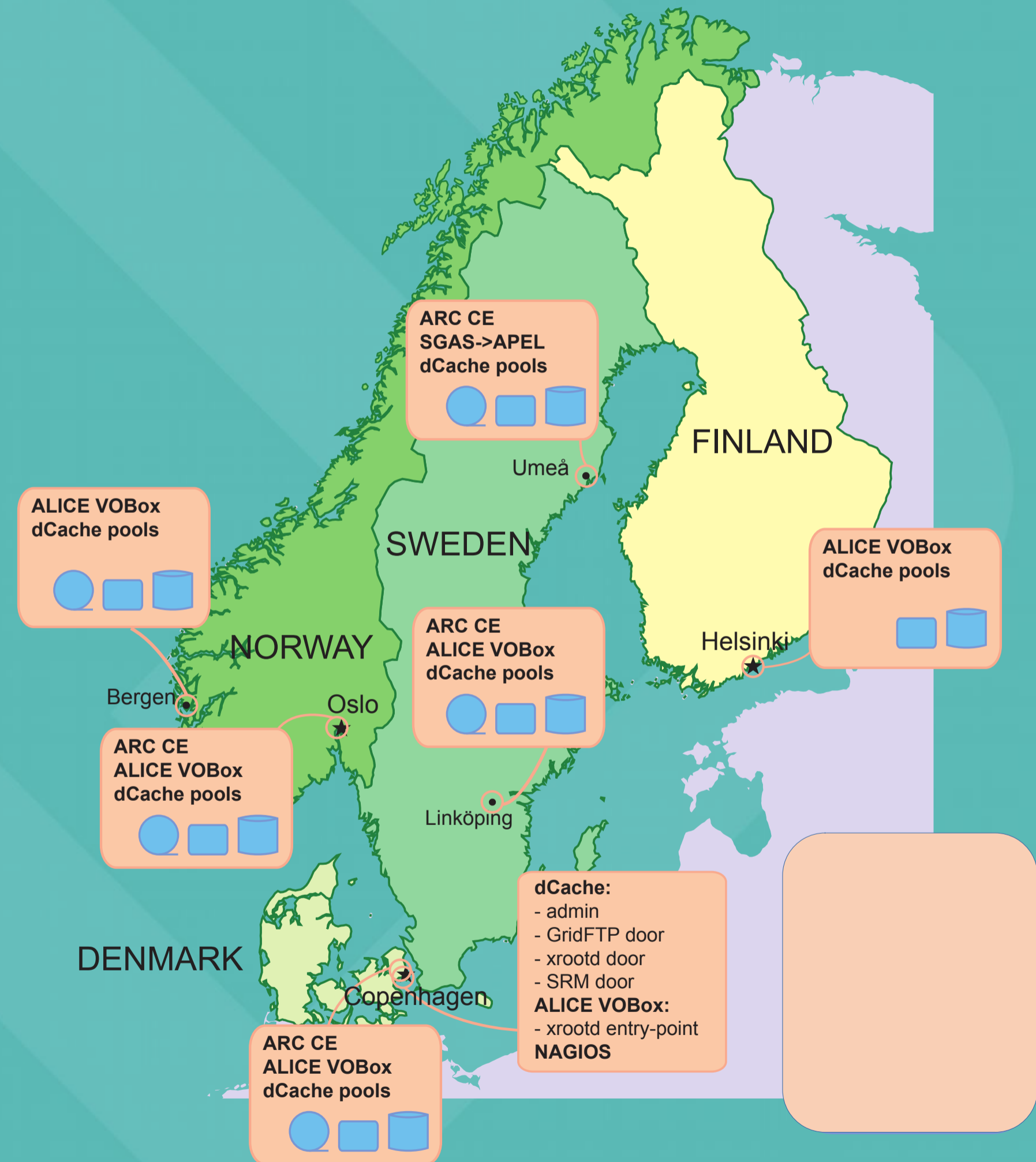
- Distributed dCache
  - Common namespace and access methods (WebDAV, xrootd etc)
  - Storage servers (“pools”) distributed
    - Some with tape libraries behind
    - Tape for less frequently used data, 3-4x cost advantage over disk
  - High Availability design
    - Most maintenance and hardware replacements done with zero user impact
- Enables us to integrate external storage
  - Slovenian, Swedish, and Swiss Tier-2 storage extends our Tier-1
    - More value to researchers for very little marginal cost to us (savings for them)





# Services: Storage

- 30 PB disk installed
  - Tier-1 pledged disk +
  - Swedish T2 +
  - Swiss (Bern) T2 +
  - Slovenian T2
    - (Including 3PB temporary commitment due to war)
- 19 PB tape installed
  - All in the Nordics



# Storage operations

- Local site admins maintain hardware, filesystem, operating system, networking, kernel tuning
  - Provides one unprivileged account with lots of storage to the central ops team
- Central ops team runs dCache pools
  - Install java + dCache
  - Configure, upgrade, restart dCache
- Investigating issues sometimes takes cooperation
  - Pool shutdown (central ops notice) due to IO error (investigation) because of raid controller issue (local ops fix)





# Pool commissioning

- Verify ssh connectivity
- Install pool, update central hosts firewalls, add to monitoring systems, etc
- Fill pool with cached copies of data appropriate for its intended use (secondary copies of ATLAS data)
  - Filling to make sure that the pool doesn't break or drop too much in performance when full – this is also a useful for all other storages
- Test reading with migration module too
- Add mixed loads and find limits
  - When does IO-load ramp up, will writes starve reads or vice versa



# Ansible

- We manage the software and configuration with Ansible
- The headnodes and all other systems will not be covered here
- Our name is “tarpool”, because it installs dCache from the tar distribution
  - As opposed from the previous setup when local admins would install it from deb or rpm
- Production and preproduction differentiated by inventory, same roles and playbooks





# Tarpool inventory

- Hosts

- Which pool nodes exists and their grouping

```
[tarpools_izum]
dcnd01.vega.izum.si
dcnd02.vega.izum.si
dcnd03.vega.izum.si
dcnd04.vega.izum.si
dcnd05.vega.izum.si
dcnd06.vega.izum.si
dcnd07.vega.izum.si
dcnd08.vega.izum.si
dcnd09.vega.izum.si
dcnd10.vega.izum.si
dcnd11.vega.izum.si
```



# Tarpool inventory

- Groupvars, things that are site-wide
  - Like homedir, username, java memory settings

```
nt1-ansible/environments/production/group_vars$ cat tarpools_csc.yml
dcache_memory_heap:      6144m
dcache_memory_direct:    4096m
pool_max_active_movers_regular_queue: 3000
pool_user_homedir:       "/var/lib/{{ pool_remote_user }}"
```





# Tarpool inventory

- `host_vars`, per pool node settings
  - List of pools defining domain, name, size, path
  - Overrides to `group_vars` when not all pools are identical at a site

```
nt1-ansible/environments/production/host_vars$ cat npool001.ijs.si.yml
pool_user_homedir:      "/home/dcache"
dcache_memory_heap:    "8192m"

dcache_poolinfo:
  pool1:
    domain: "npool001_ijs_si_095Domain"
    poolname: "ijs_si_095"
    poolsize: "300T"
    poolpath: "/cephpools/pool095"
```



# Tarpool playbook

- Install and upgrade software
  - Java and dCache versions pinned with URLs to tarballs
- Configure
  - Apply current ansible configuration to pool config files
- Settings
  - Set settings through the admin interface for live adjustments (max movers, etc)
- Start, stop, restart
  - Starts, stops, and restarts the pools
- Create
  - Creates pools according to inventory rules





# Tarpool operations

- OS and firmware upgrade sometimes need reboots
  - Local site admins are responsible for driving this process
  - Ops team needs to set readonly ahead of time and shut down dCache on the pool
  - We provide a trigger file to tell that dCache has been intentionally shut down
  - Some sites use this for automatic reboots into new kernels
    - `if(dcachelogflagfile and reboot_required) then reboot`
  - Less urgent stuff can be co-ordinated with other downtimes



# Grafana graphs

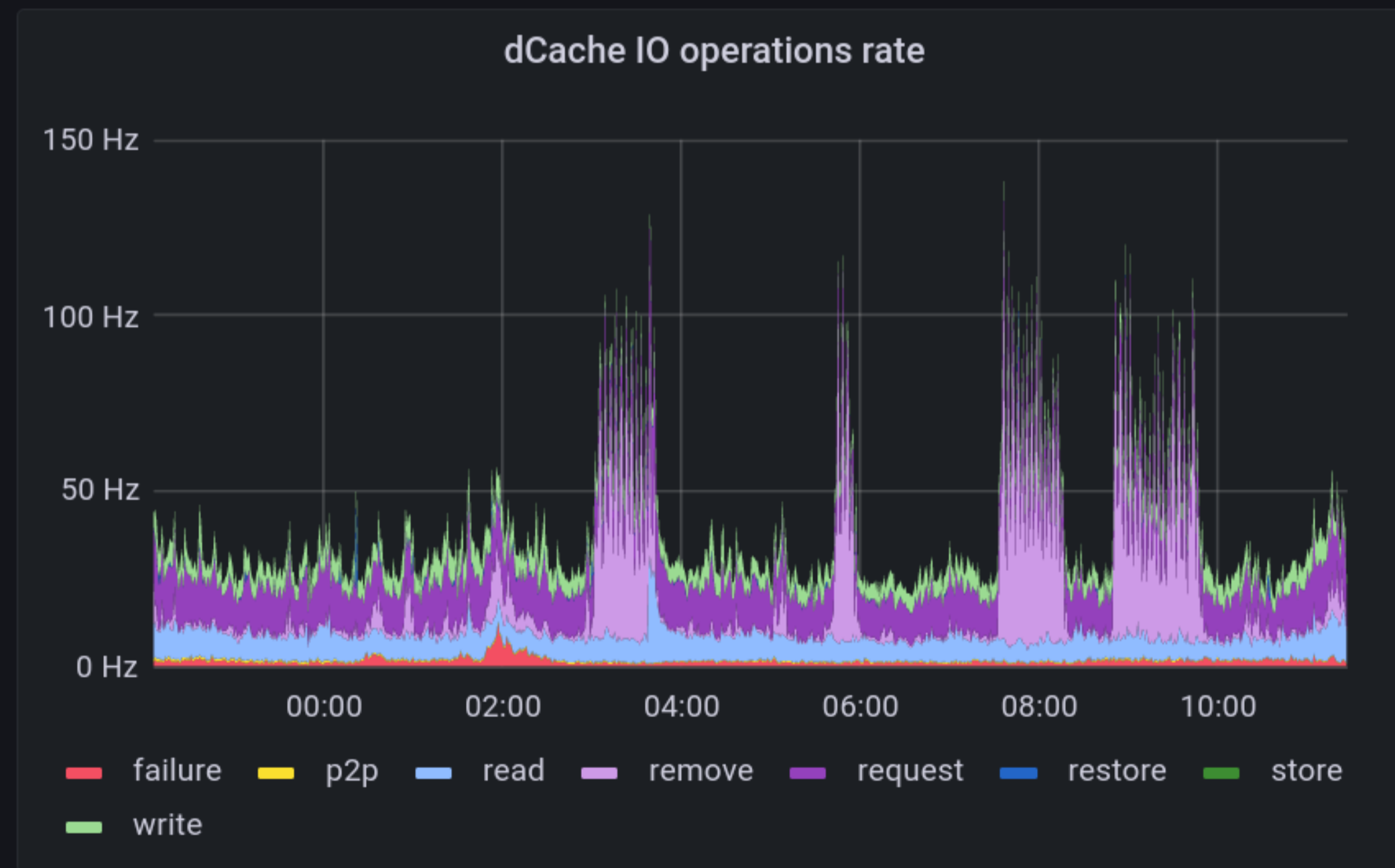
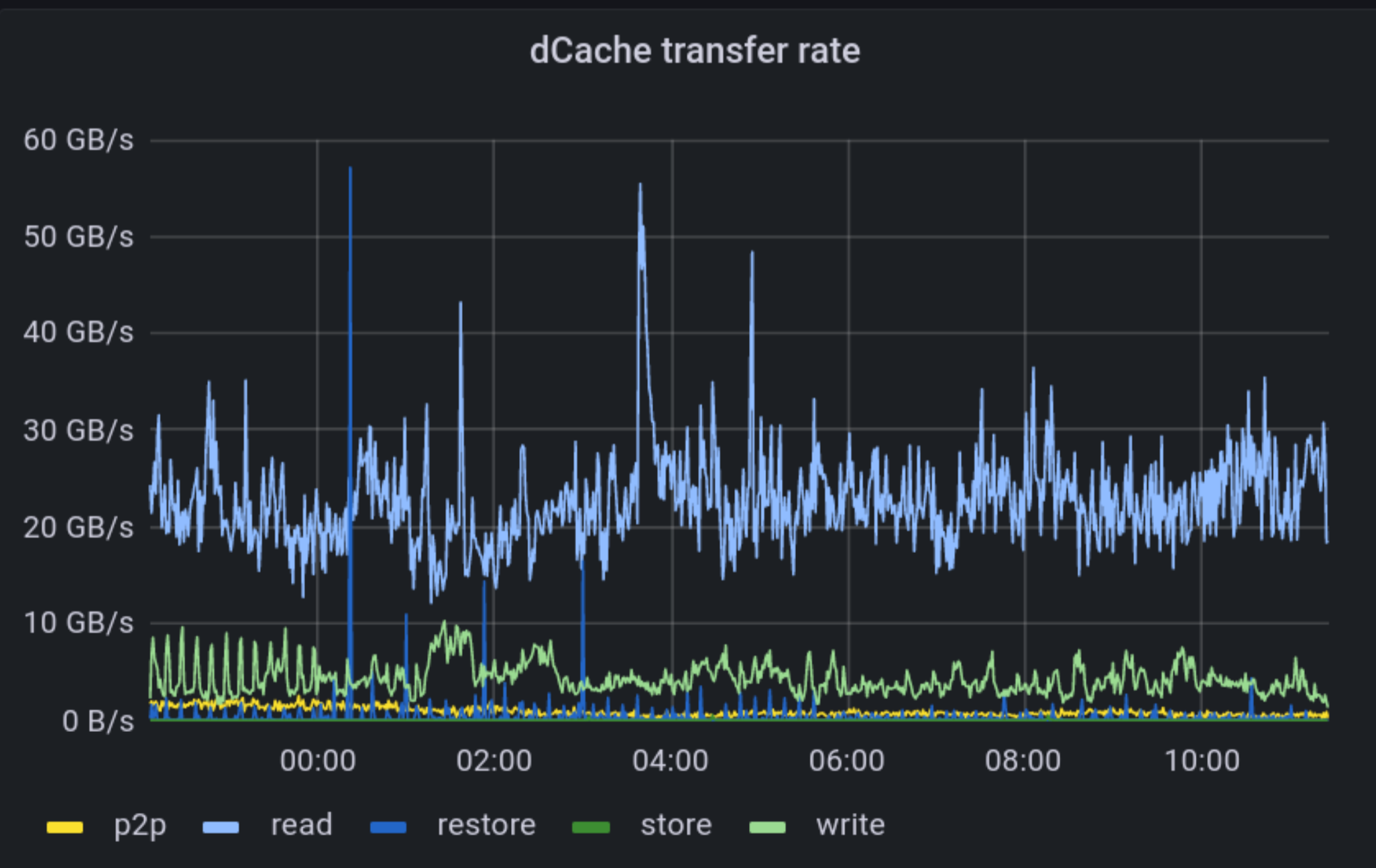
- We're in the middle of moving from ganglia to a new monitoring stack: victora metrics, grafana, prometheus, etc
- These graphs still have a few issues, see if you can find them all!



# Top level graph

- Throughput rates, and operations

## ▼ dCache Statistics

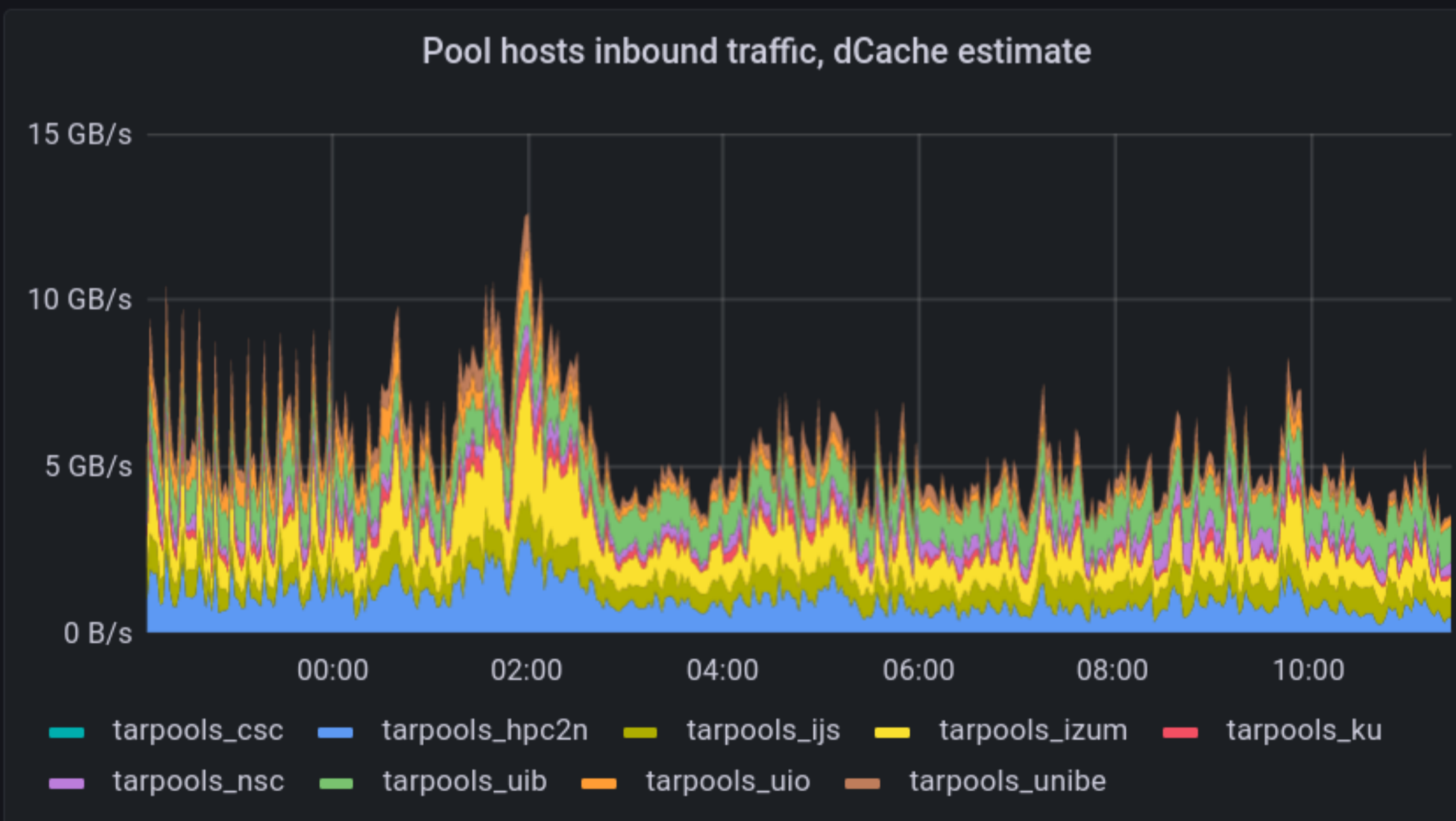
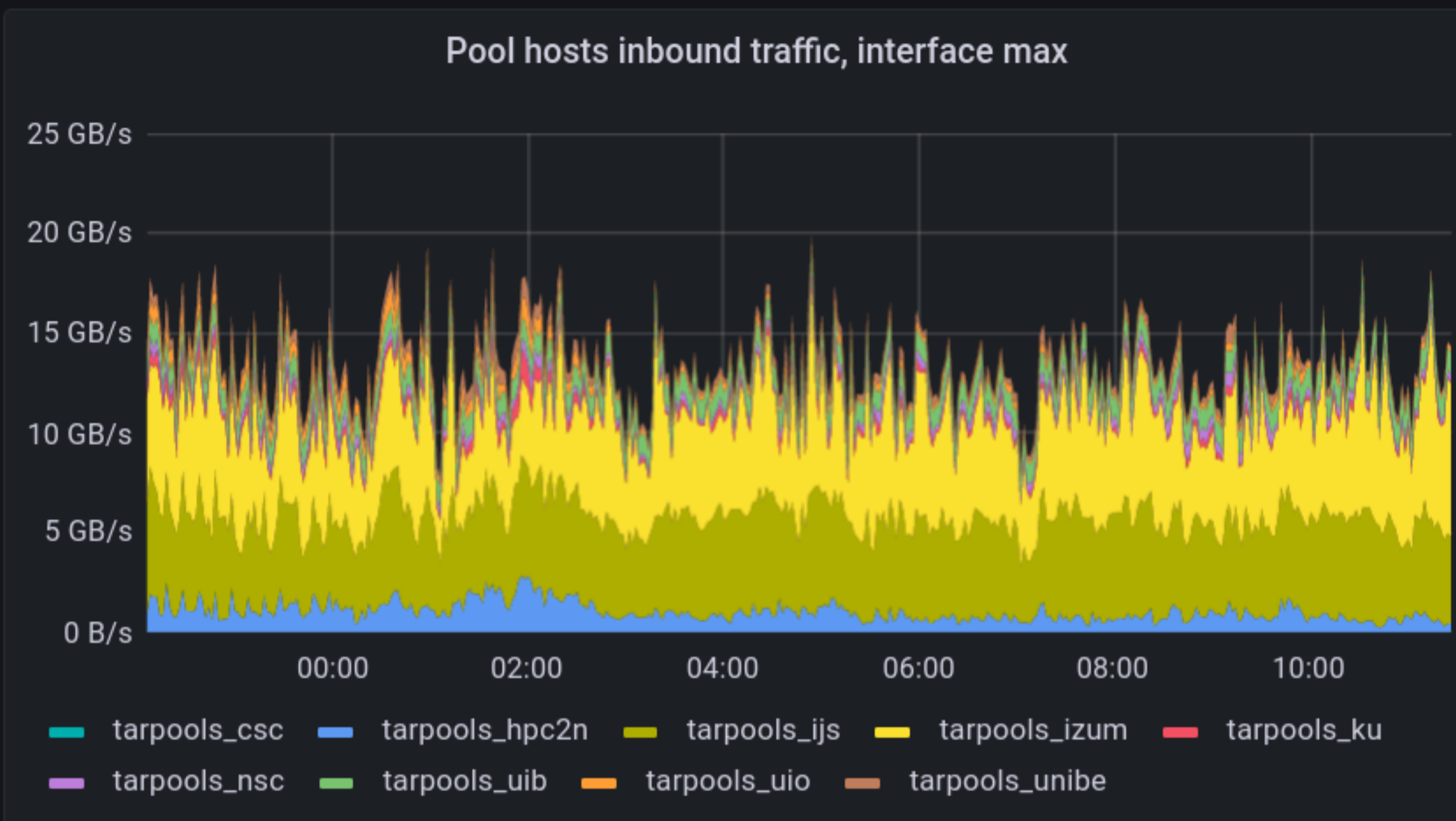




# Traffic

- Inbound traffic split by pool
- Two views, OS probes of network traffic and dCache billing logs
- Writes to us

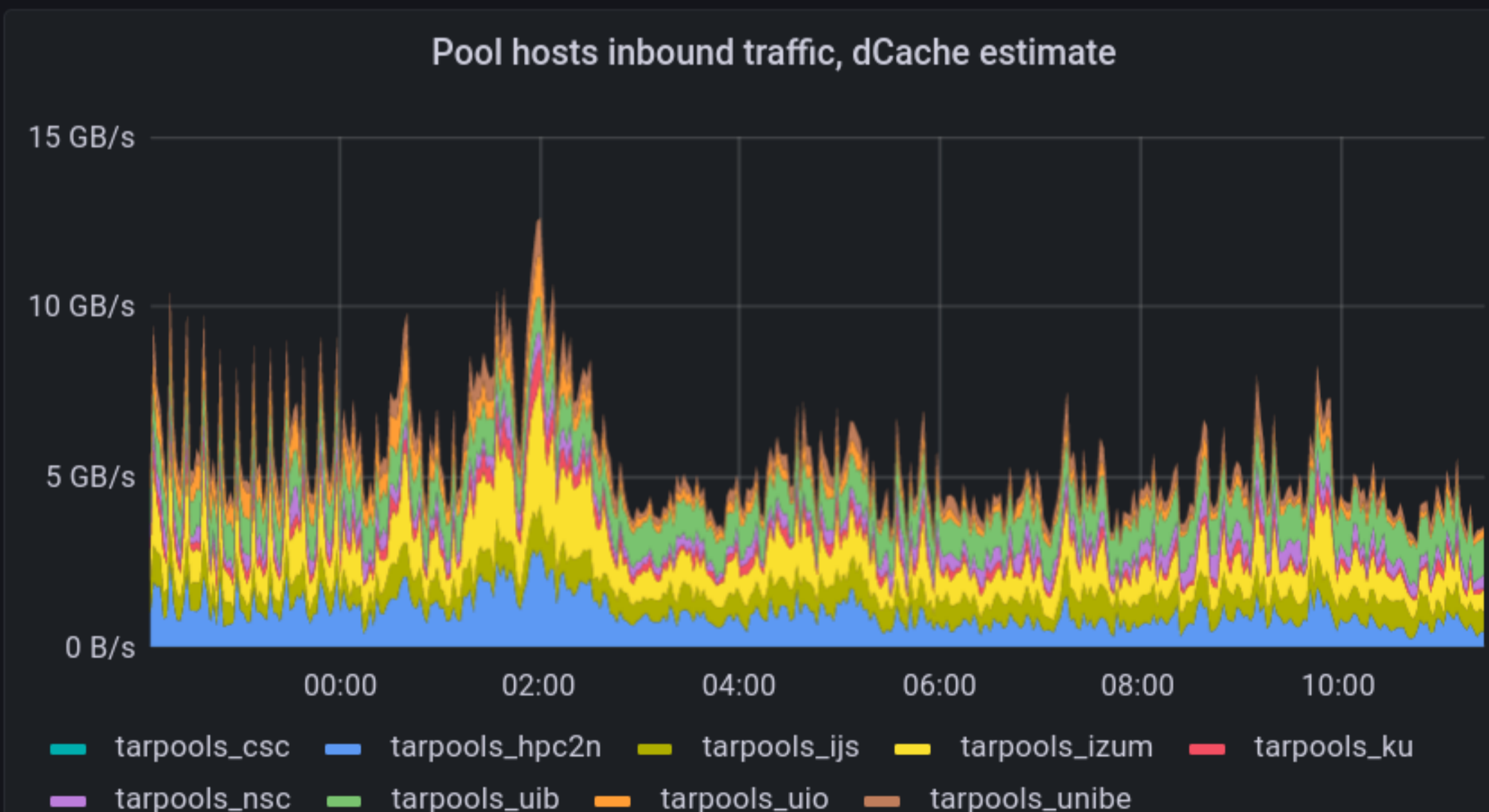
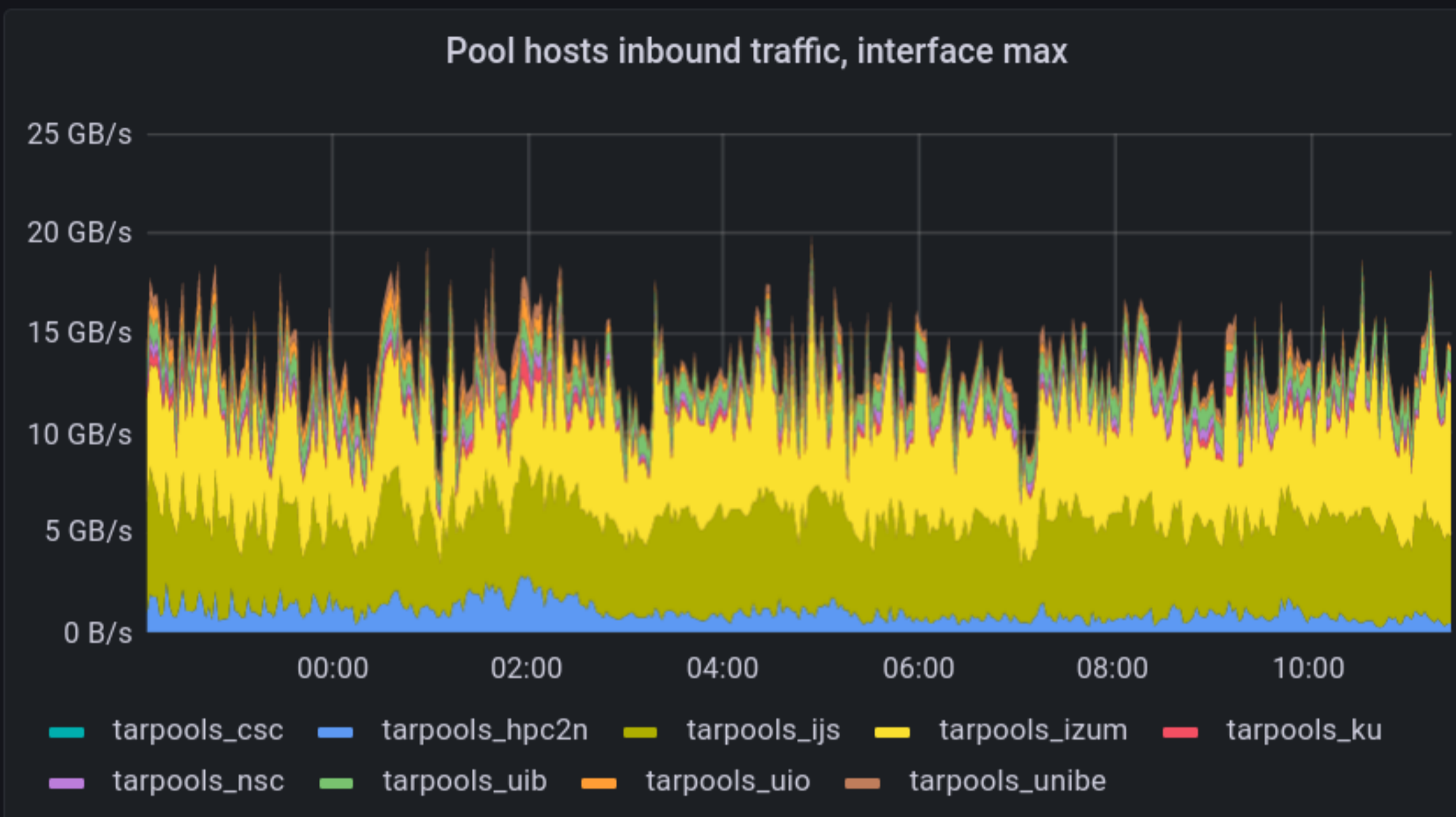
## Aggregated Pool Traffic



# Traffic

- Inbound traffic split by pool
- Two views, OS probes of network traffic and dCache billing logs
- Writes to us
- But also CephFS traffic for reads in Slovenian pools in the top graph

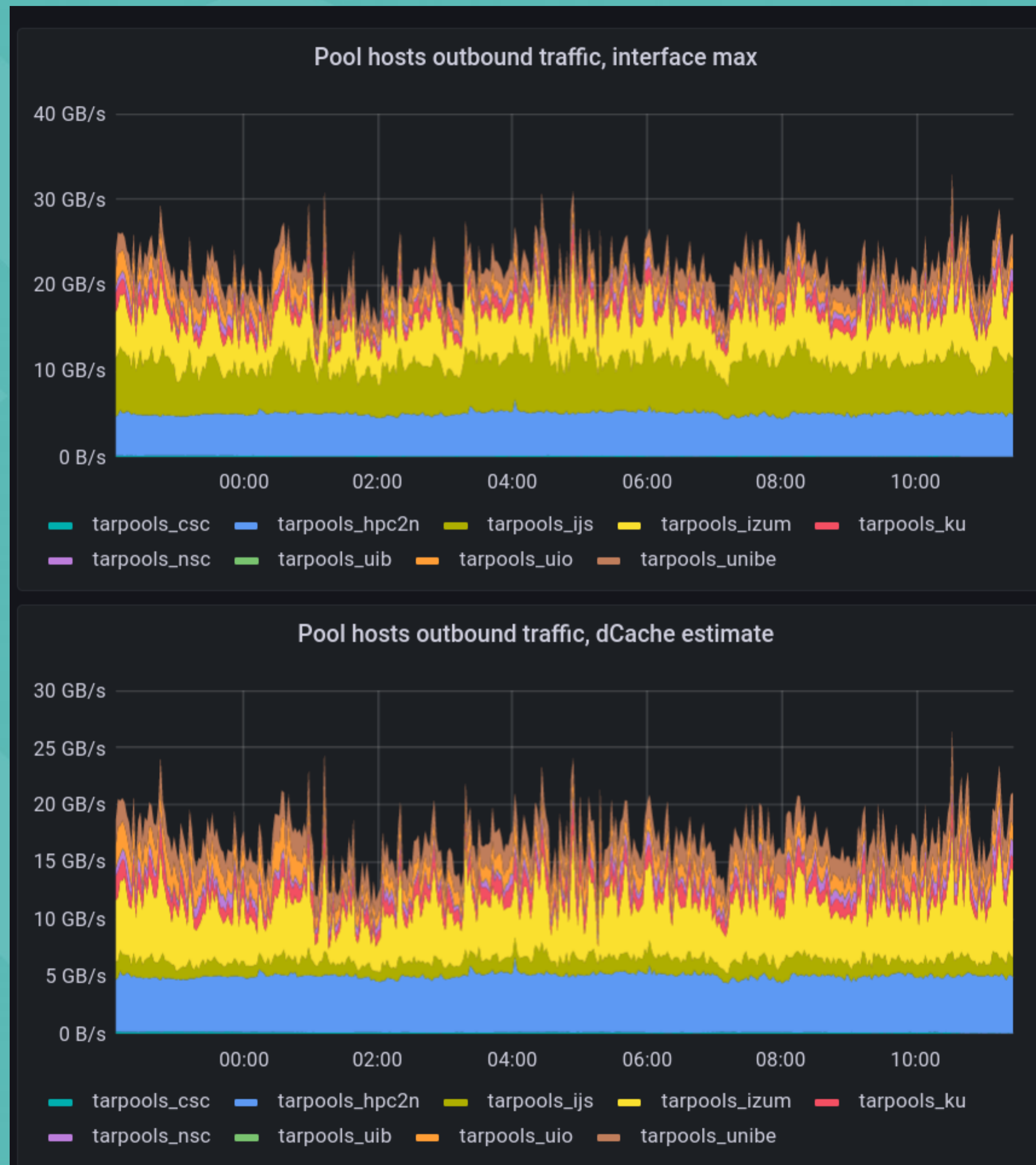
## Aggregated Pool Traffic





# Traffic

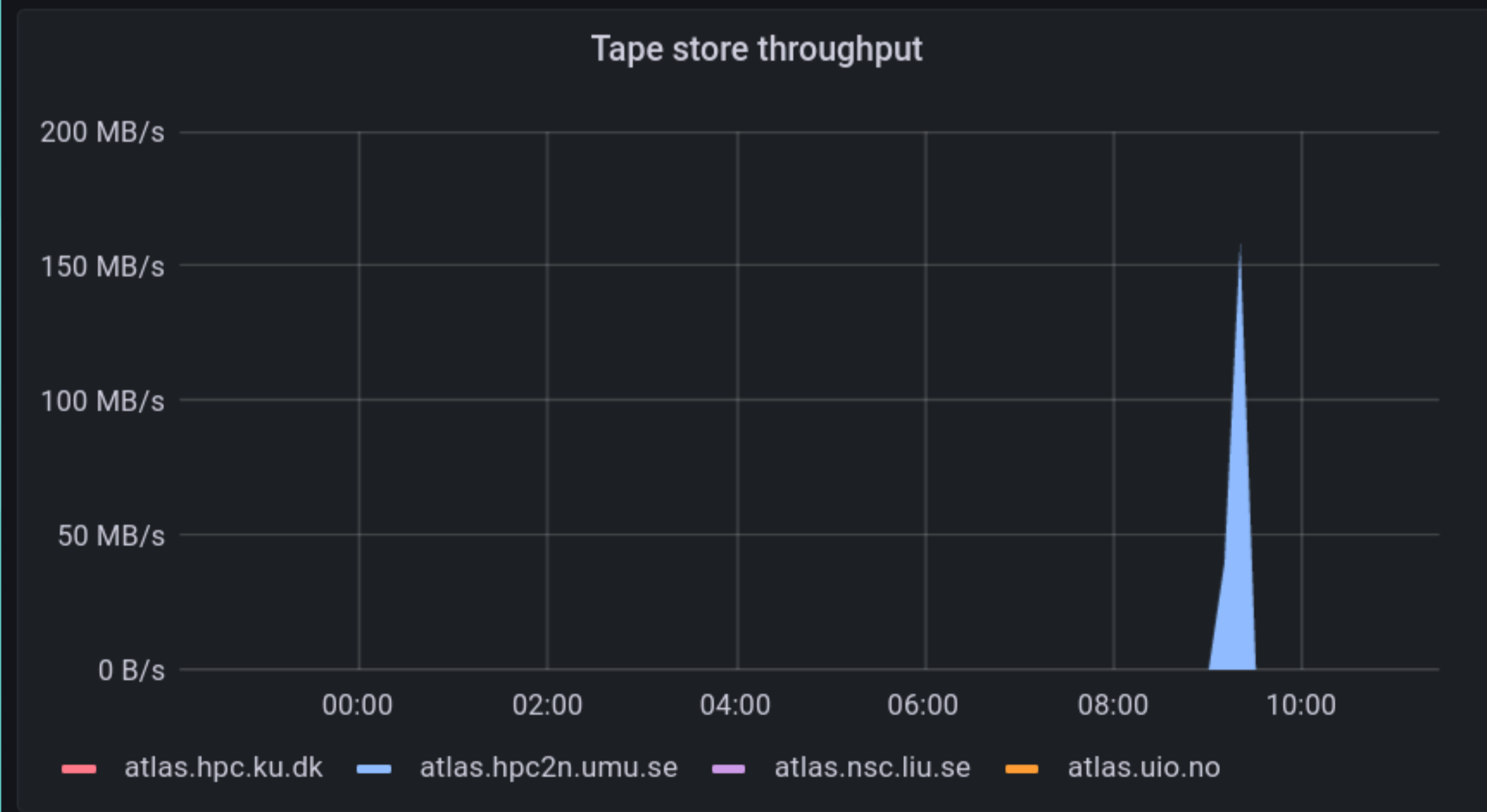
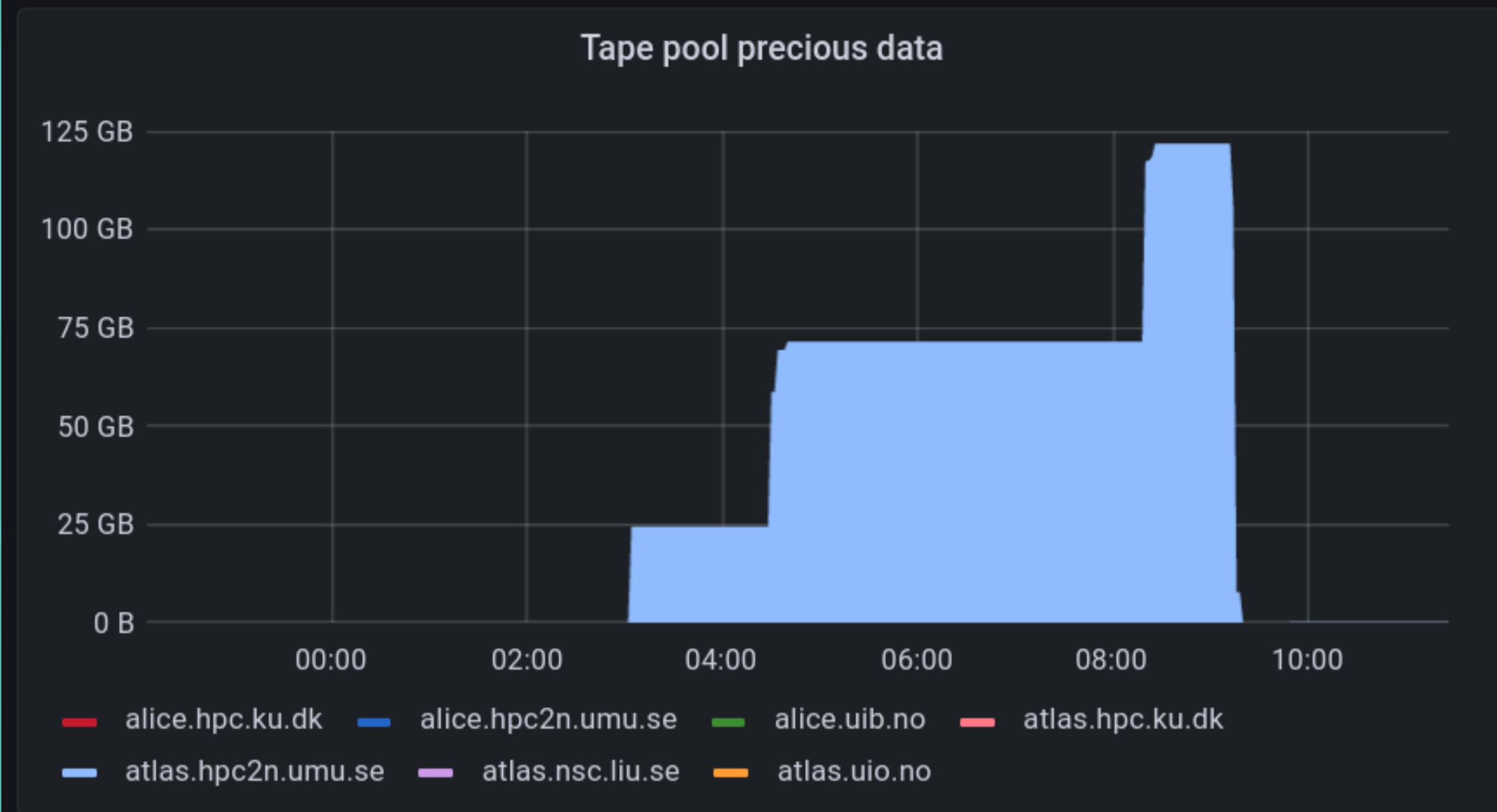
- Outbound traffic split by pool
- Two views, OS probes of network traffic and dCache billing logs
- Reads from us
- But also extra CephFS traffic
- HPC2N network limit





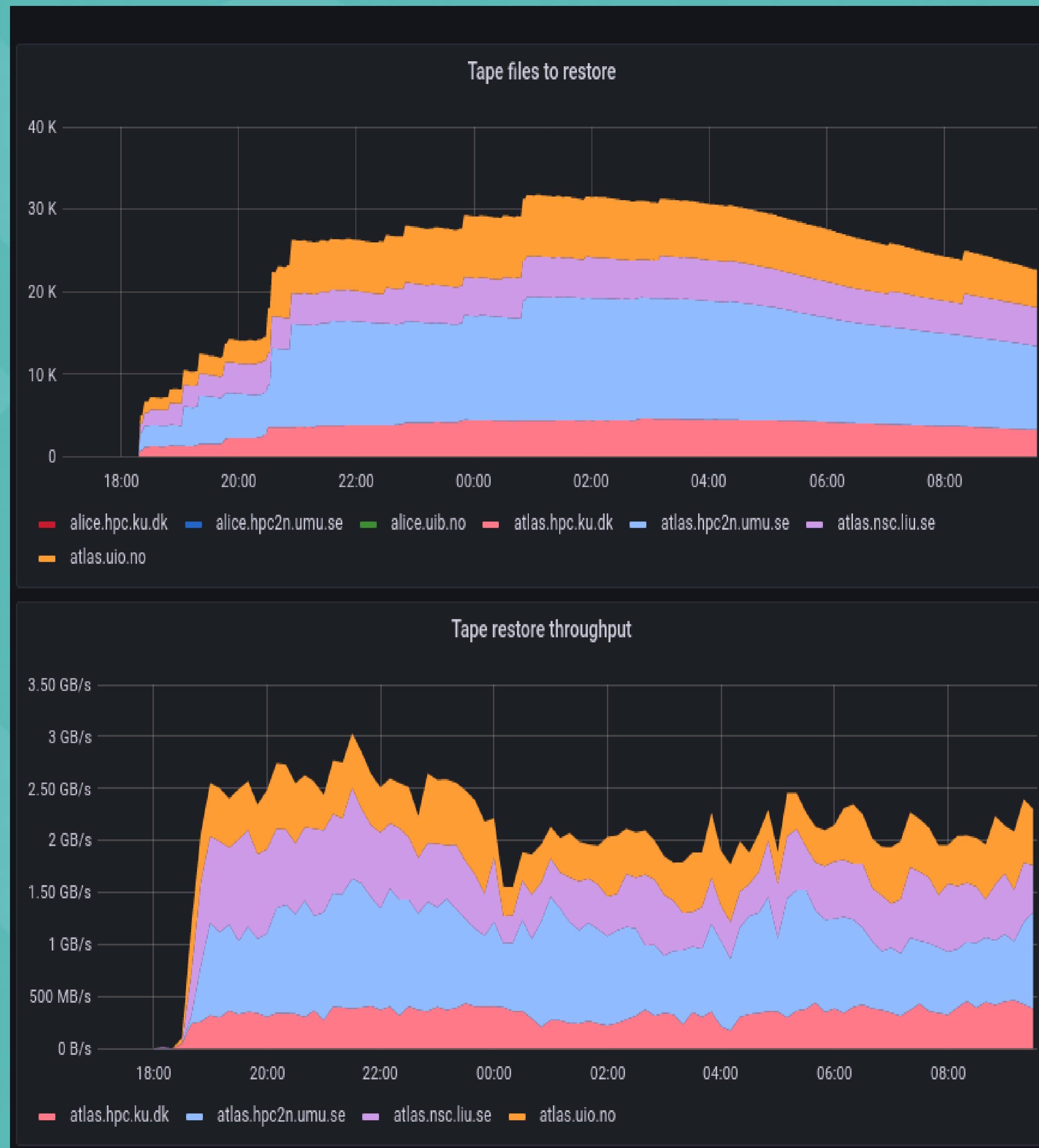
# Tape store

- Pretty boring period
- ENDIT waits a while for batching tape writes
  - (size > x GB) or (time > n seconds)
  - Tuned by local tape experts for their tape library
  - 6h in this case



# Tape write

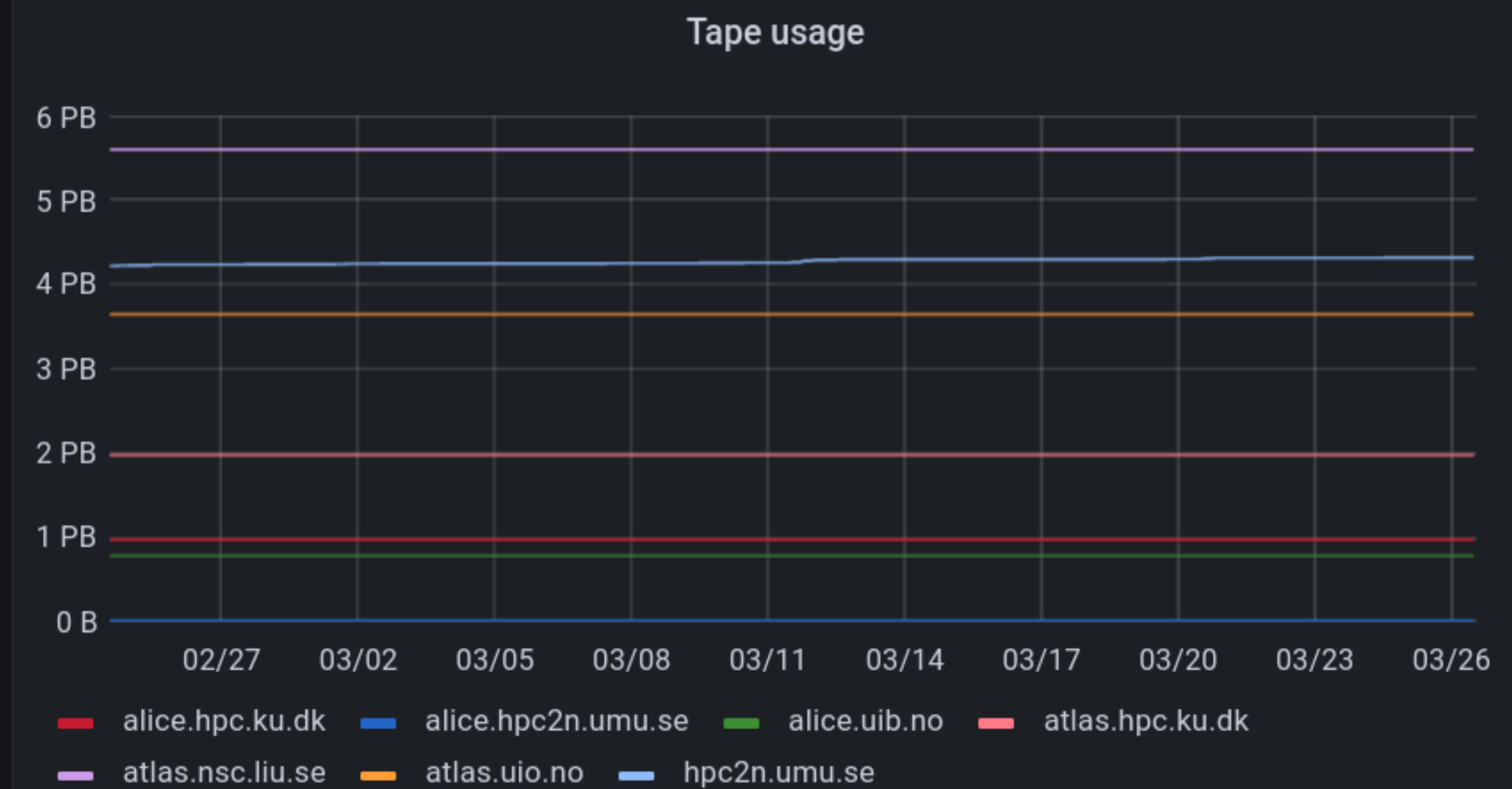
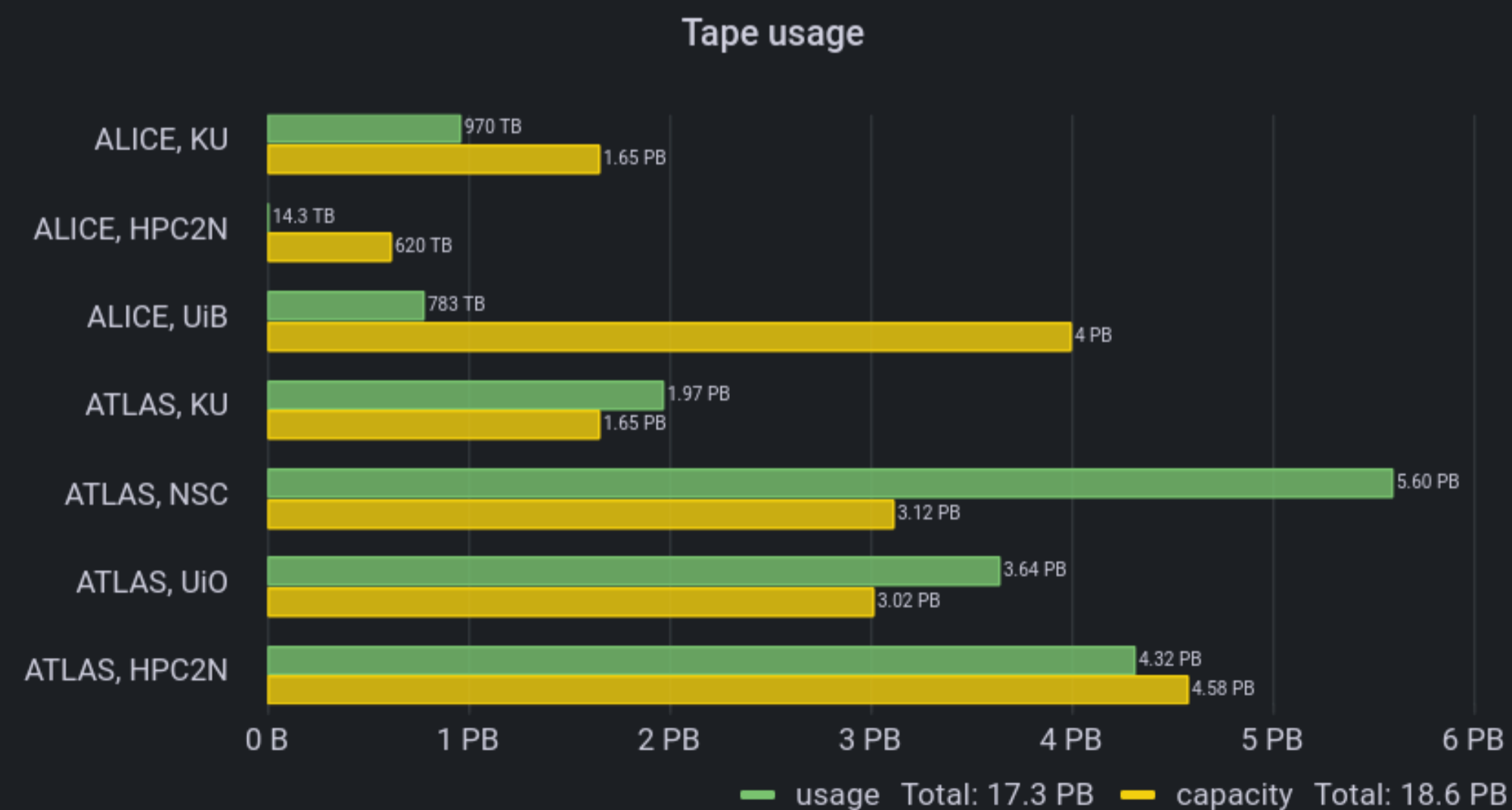
- Less boring period
- Some of our tape systems are faster than others
- Number of tape drives used depends on queue size and local max



# Tape overview

- Used vs installed, and long-term fill
  - Used > installed means borrowed capacity

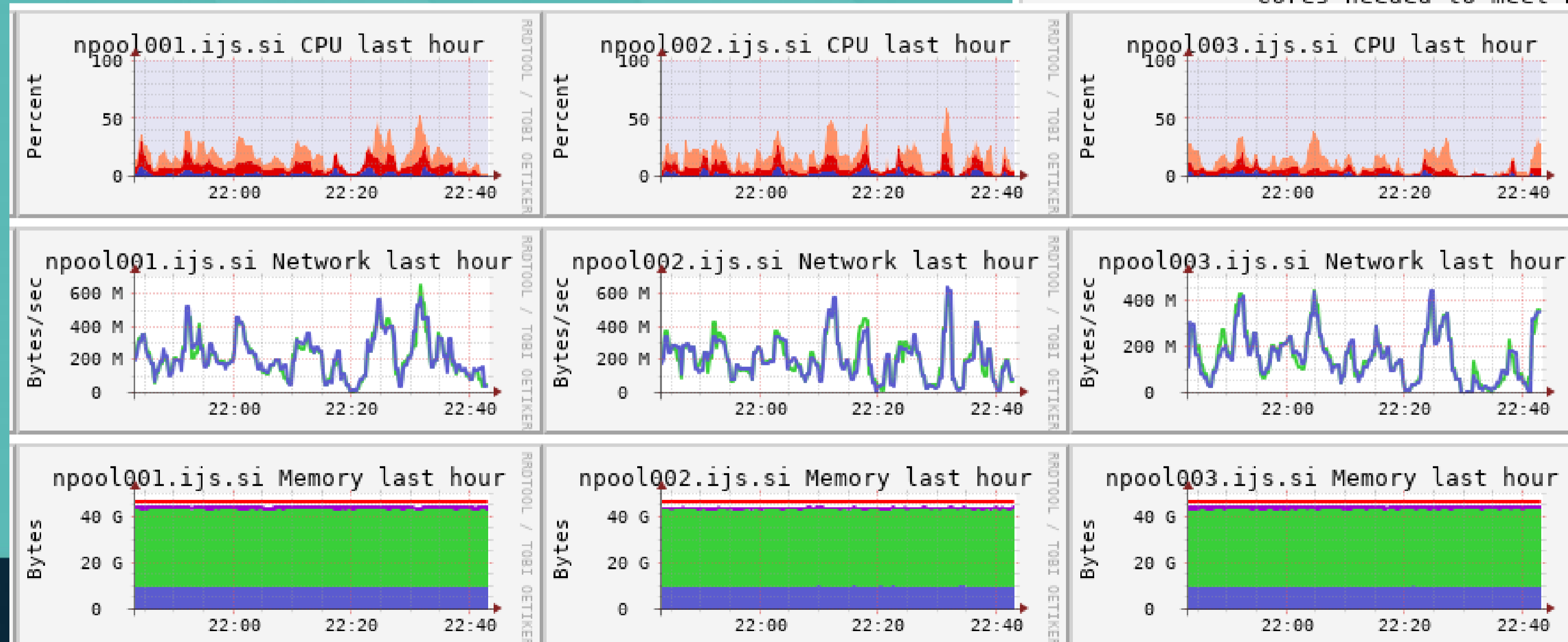
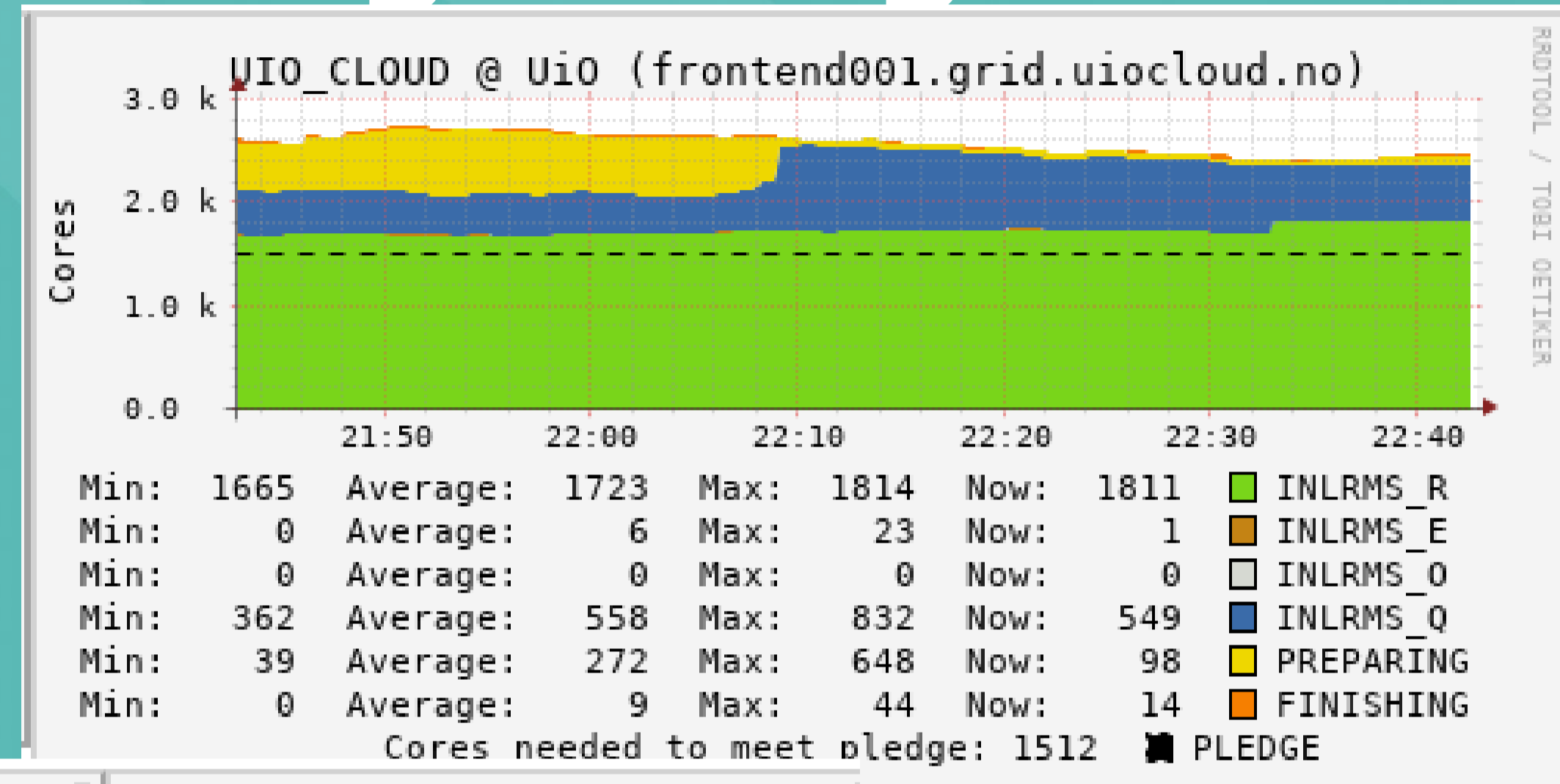
## ✓ Tape Usage Statistics





# Legacy monitoring: Ganglia

- Migrating to grafana
- ARC monitoring:
- dCache pool hosts:



# dCache admin

- We create a site pool group for all sites

PoolGroup	Total Space/MB	Free Space/MB	Precious Space/MB	Layout (precious/sticky/cached free)
<a href="#">alice_disk</a>	6622773860	2298412455	0	
<a href="#">alice_tape_read</a>	53474644	1796	0	
<a href="#">alice_tape_write</a>	9437184	1063	0	
<a href="#">alice_tape_write_backup</a>	23359488	23354113	0	
<a href="#">atlas_disk</a>	23258294272	236630392	0	
<a href="#">atlas_disk_hpc2n</a>	4563402752	24478160	0	
<a href="#">atlas_disk_ijs</a>	4805658624	152896475	0	
<a href="#">atlas_disk_ku</a>	1071104000	2434156	0	
<a href="#">atlas_disk_lhep</a>	1838153728	16628740	0	
<a href="#">atlas_disk_nsc</a>	2547335168	15253826	0	
<a href="#">atlas_disk_uio</a>	1658880000	8346564	0	
<a href="#">atlas_disk_vega</a>	6438912000	16199029	0	
<a href="#">atlas_local_disk_uio</a>	1658880000	8346564	0	
<a href="#">atlas_tape_read</a>	66254646	6736	0	
<a href="#">atlas_tape_write</a>	13312000	2163	142	
<a href="#">atlas_tape_write_backup</a>	23040000	3004	142	
<a href="#">default</a>	30308753485	2797939800	143	
<a href="#">dteam_disk</a>	1867276288	9980201	0	
<a href="#">dteam_tape_read</a>	1048576	1037219	0	
<a href="#">dteam_tape_write</a>	1048576	183	0	
<a href="#">ops_disk</a>	16384	13717	0	
<a href="#">ska_disk</a>	1867276288	9980201	0	
<a href="#">test_disk</a>	1638924288	9241491	0	
<a href="#">test_tape_read</a>	1048576	1037219	0	
<a href="#">test_tape_write</a>	1048576	183	0	



# dCache admin

- Extensive usage of wildcards and site poolgroups

```
[piggy] (local) admin > \s hpc2n_umu_se_*/atlas_disk rep ls -s
```

```
[piggy] (local) admin > \s  
vega_izum_si_039,vega_izum_si_040,vega_izum_si_041,vega_izum_si_042 migration move -  
target=pgroup atlas_disk_ijs
```

```
[piggy] (PoolManager@piggyDomain) admin > psu set pool hpc_ku_dk_* rdonly
```







Questions?



# References

- Pool hardware guidelines for our site:
  - [https://wiki.neic.no/wiki/DCache\\_Pool\\_Hardware](https://wiki.neic.no/wiki/DCache_Pool_Hardware)
- Pool installation guide for site admins:
  - [https://wiki.neic.no/wiki/DCache\\_Pool\\_installation](https://wiki.neic.no/wiki/DCache_Pool_installation)
- Linux tuning for dCache pools:
  - [https://wiki.neic.no/wiki/Operations\\_Tuning\\_Linux](https://wiki.neic.no/wiki/Operations_Tuning_Linux)

