Science and Technology Facilities Council

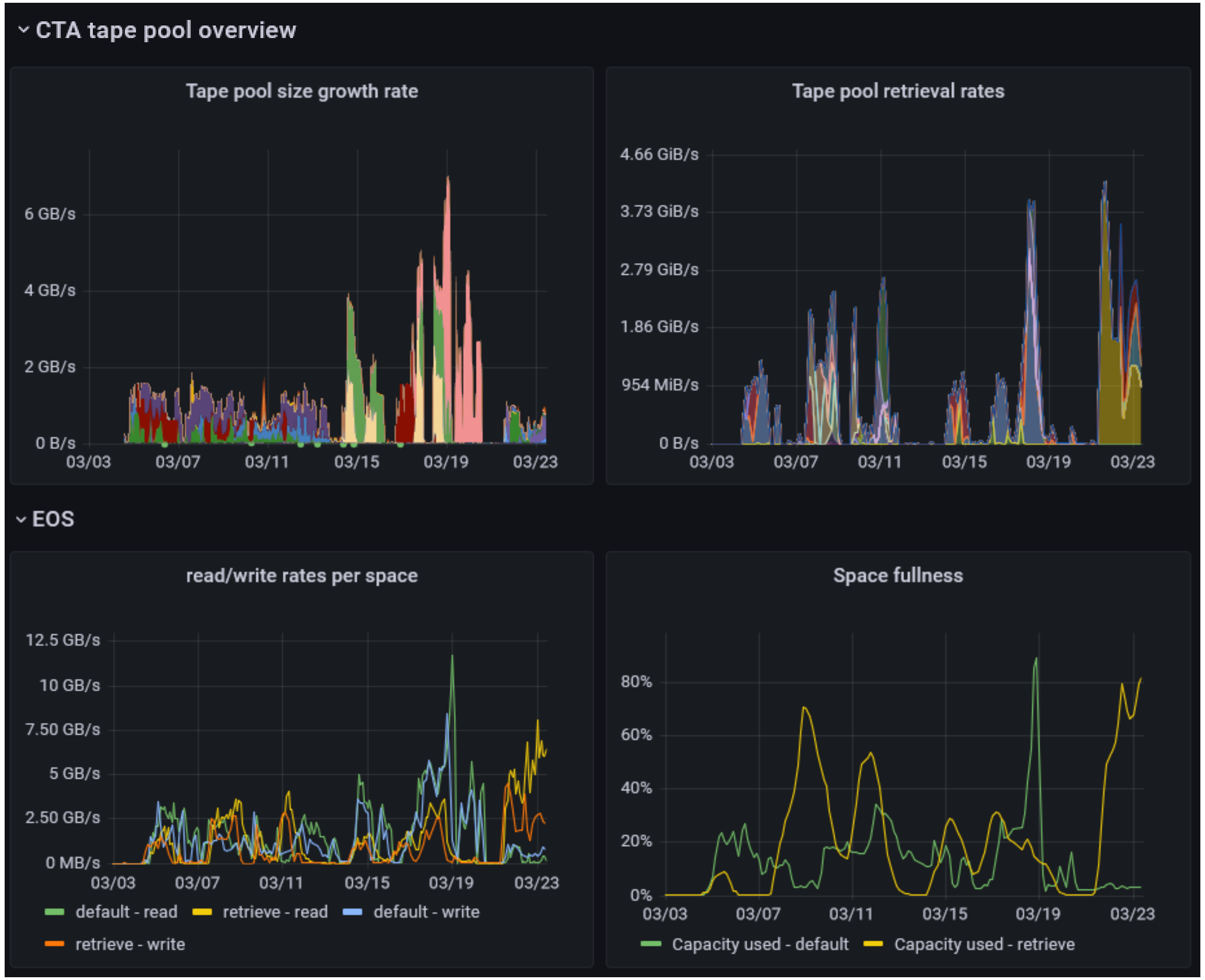
RAL Antares Update

Tom Byrne, George Patargias

23rd March 2022

Recent progress highlights – GridPP47

1. Participated in the 2021 LHC data challenge
2. Planned and tested migration from RAL WLCG CASTOR to EOS+CTA
3. Finalised EOS+CTA setup at RAL and rebuilt the production instance at full scale
4. Migrated from CASTOR to EOS+CTA
5. Antares has been in production for 19 days
6. Currently participating in another LHC data challenge

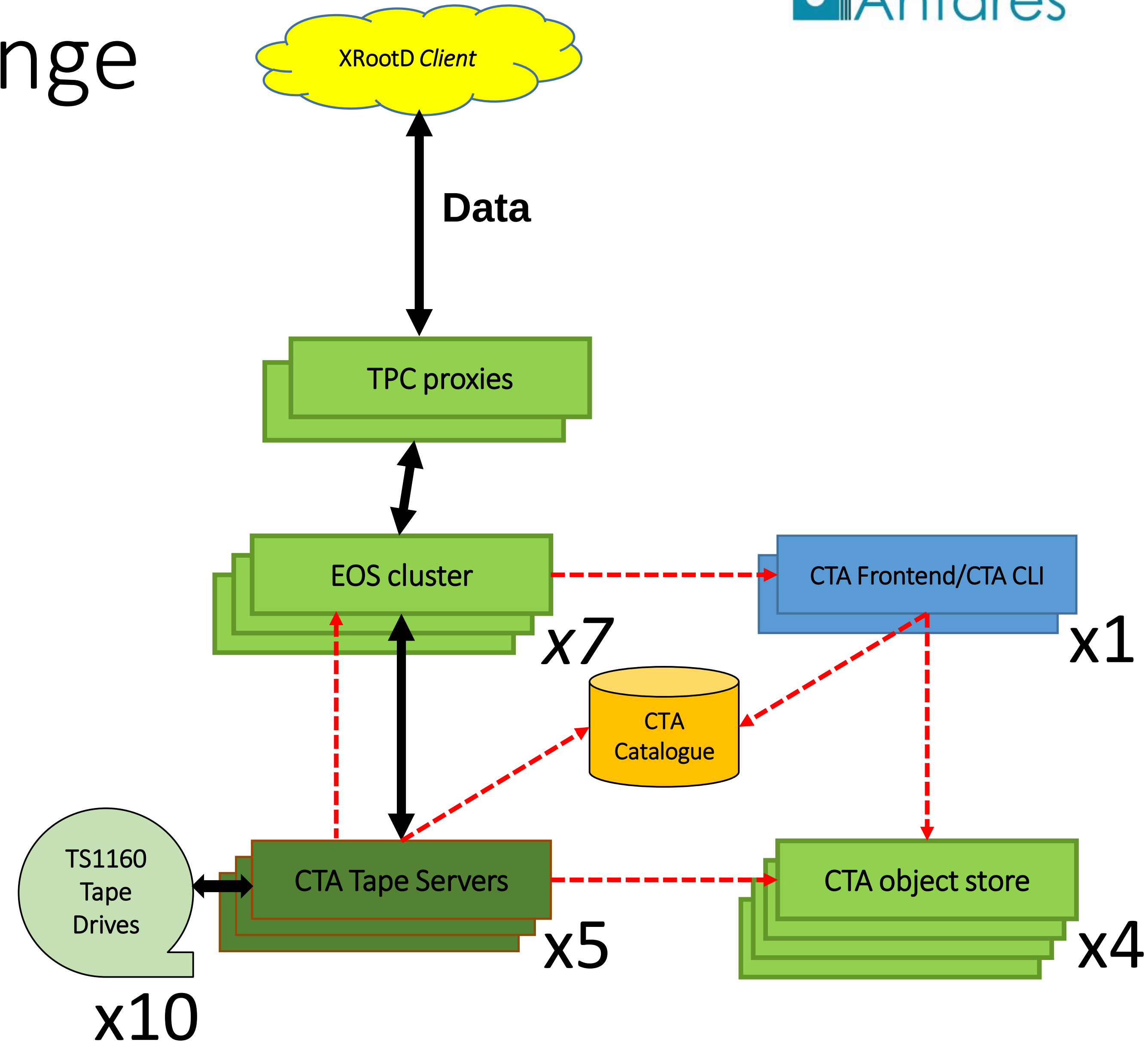


First 19 days of production

Science and Technology Facilities Council

2021 LHC data challenge

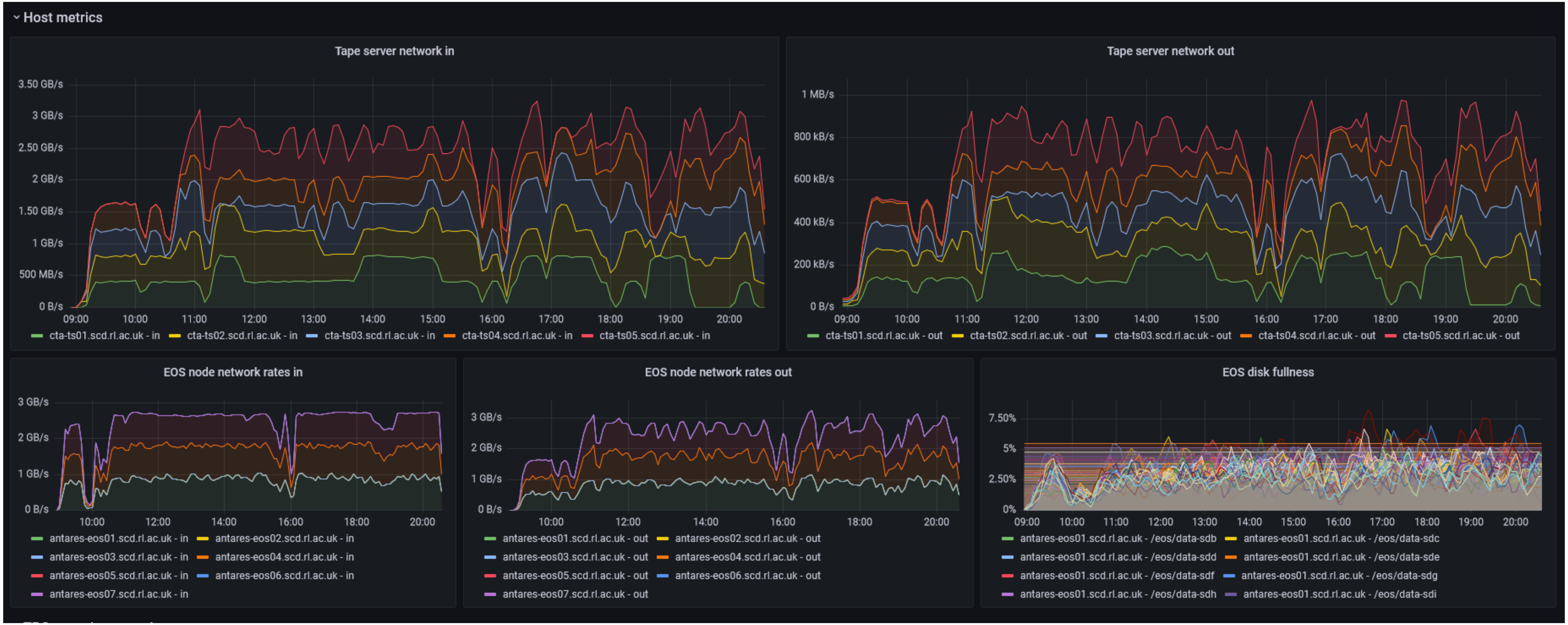
- The first of several data challenges to ensure custodial data storage systems are ready for Run 3 rates
  - Was a good opportunity to validate CTA performance with these tests
- Deployed the largest EOS+CTA stack we had ever run for this DC
- Lots of learning done – a very valuable experience for a brand new system



Tier-1 Spectra Logic Tfinity tape library

2021 LHC data challenge – Day 1 – ATLAS and LHCb Archiving

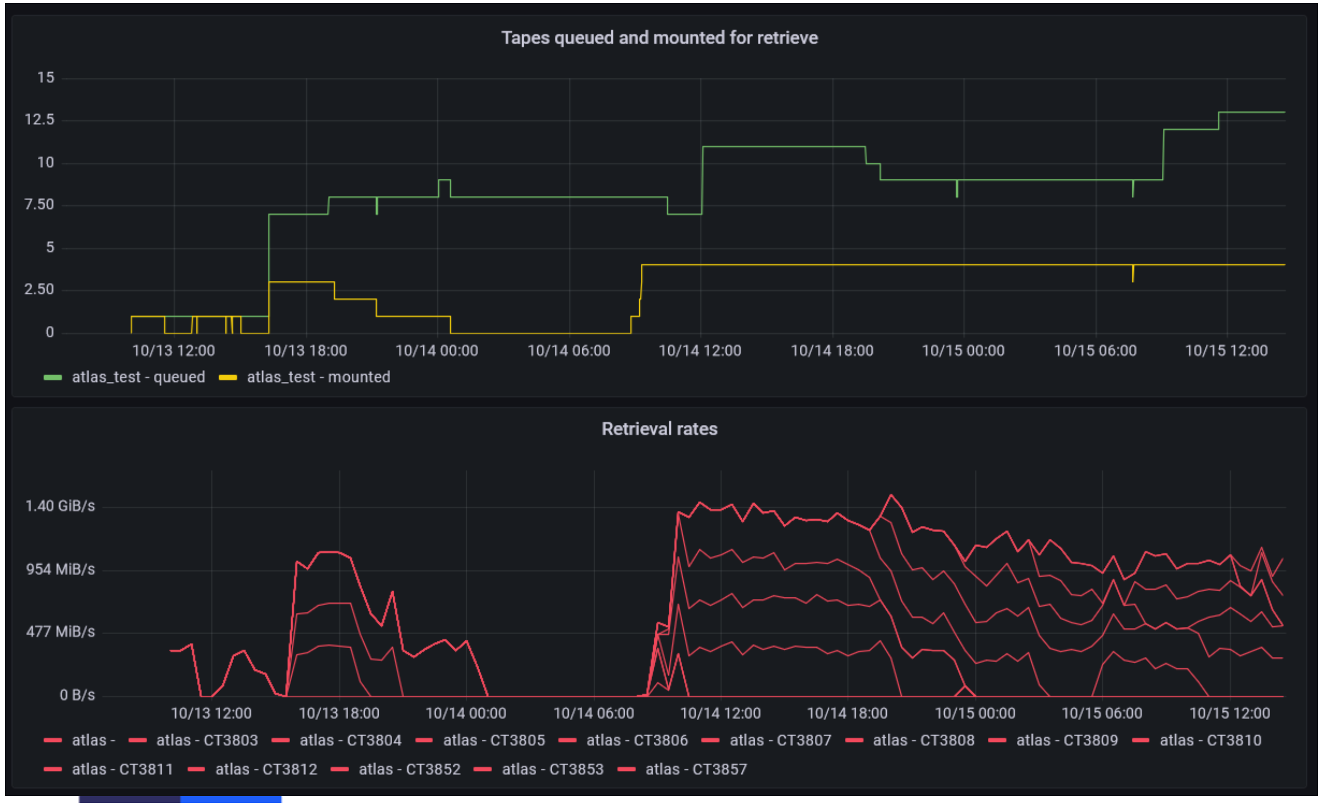
0900 – 2100 11th October

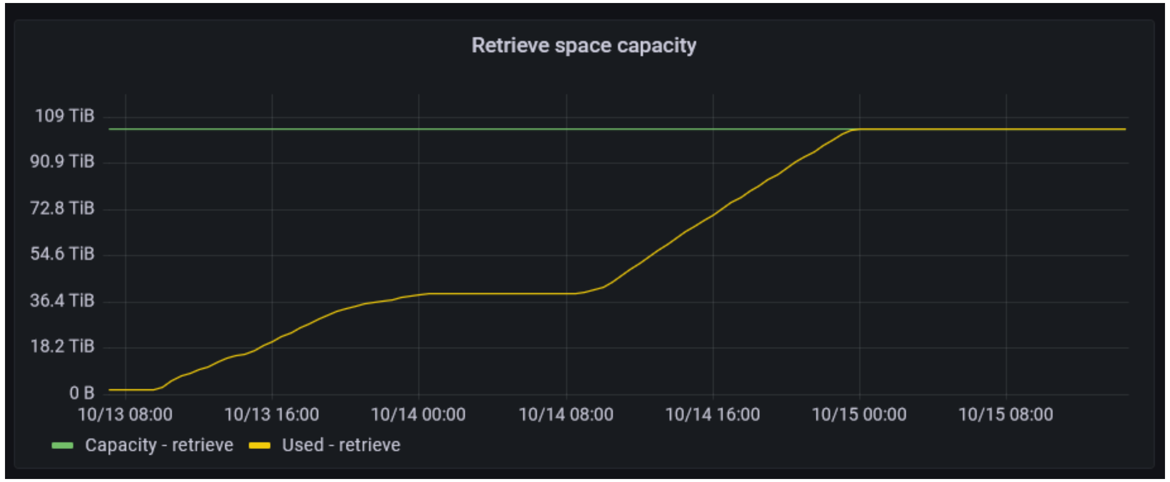

Host metrics
Tape server network in
3.50 GB/s
3 GB/s
2.50 GB/s
2 GB/s
1.50 GB/s
1 GB/s
500 MB/s
0 B/s
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
cta-ts01.scd.rl.ac.uk - in
cta-ts02.scd.rl.ac.uk - in
cta-ts03.scd.rl.ac.uk - in
cta-ts04.scd.rl.ac.uk - in
cta-ts05.scd.rl.ac.uk - in
Tape server network out
1 MB/s
800 kB/s
600 kB/s
400 kB/s
200 kB/s
0 B/s
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
cta-ts01.scd.rl.ac.uk - out
cta-ts02.scd.rl.ac.uk - out
cta-ts03.scd.rl.ac.uk - out
cta-ts04.scd.rl.ac.uk - out
cta-ts05.scd.rl.ac.uk - out
EOS node network rates in
3 GB/s
2 GB/s
1 GB/s
0 B/s
10:00
12:00
14:00
16:00
18:00
20:00
antares-eos01.scd.rl.ac.uk - in
antares-eos02.scd.rl.ac.uk - in
antares-eos03.scd.rl.ac.uk - in
antares-eos04.scd.rl.ac.uk - in
antares-eos05.scd.rl.ac.uk - in
antares-eos06.scd.rl.ac.uk - in
antares-eos07.scd.rl.ac.uk - in
EOS node network rates out
3 GB/s
2 GB/s
1 GB/s
0 B/s
10:00
12:00
14:00
16:00
18:00
20:00
antares-eos01.scd.rl.ac.uk - out
antares-eos02.scd.rl.ac.uk - out
antares-eos03.scd.rl.ac.uk - out
antares-eos04.scd.rl.ac.uk - out
antares-eos05.scd.rl.ac.uk - out
antares-eos06.scd.rl.ac.uk - out
antares-eos07.scd.rl.ac.uk - out
EOS disk fullness
7.50%
5%
2.50%
0%
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
antares-eos01.scd.rl.ac.uk - /eos/data-sdb
antares-eos01.scd.rl.ac.uk - /eos/data-sdc
antares-eos01.scd.rl.ac.uk - /eos/data-sdd
antares-eos01.scd.rl.ac.uk - /eos/data-sde
antares-eos01.scd.rl.ac.uk - /eos/data-sdf
antares-eos01.scd.rl.ac.uk - /eos/data-sdg
antares-eos01.scd.rl.ac.uk - /eos/data-sdh
antares-eos01.scd.rl.ac.uk - /eos/data-sdi

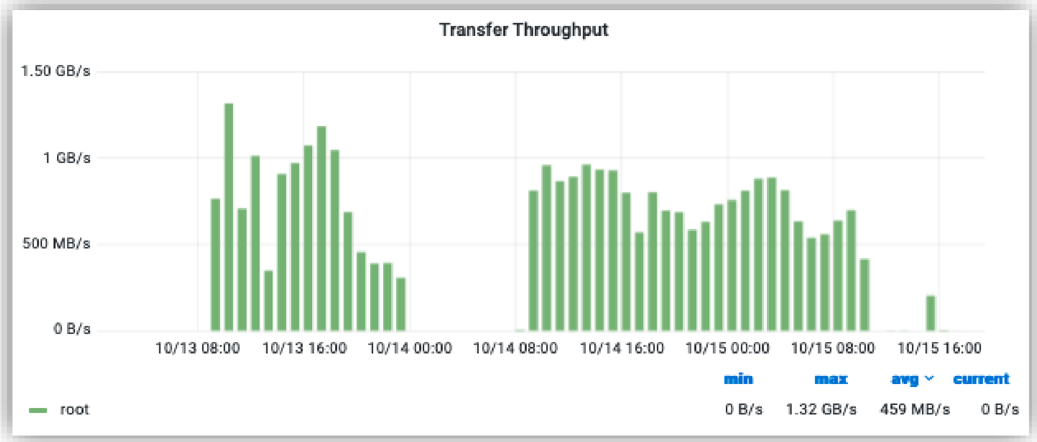
2021 LHC data challenge – Day 3/4 – Atlas recall

0800 13th – 0800 15th October

Without specific ‘xrootd.site’ name specified, FTS will not evict transferred file



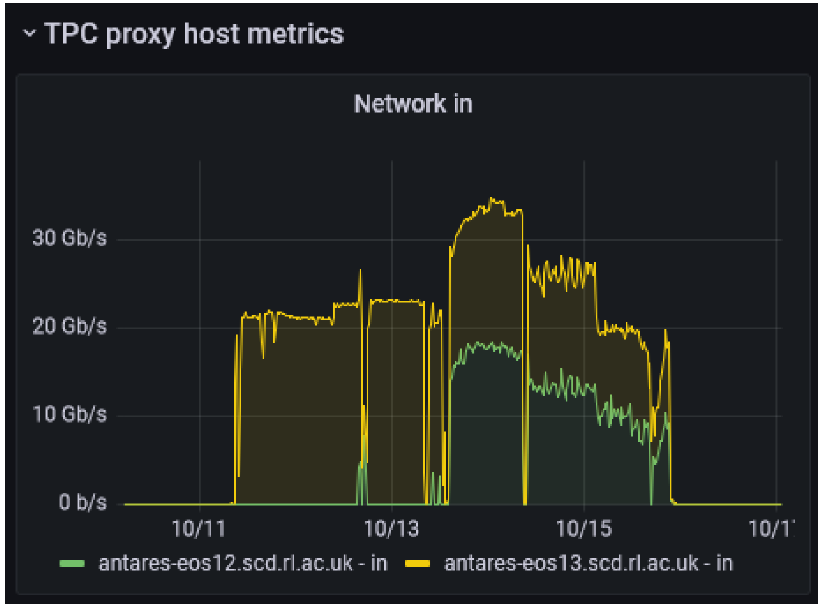




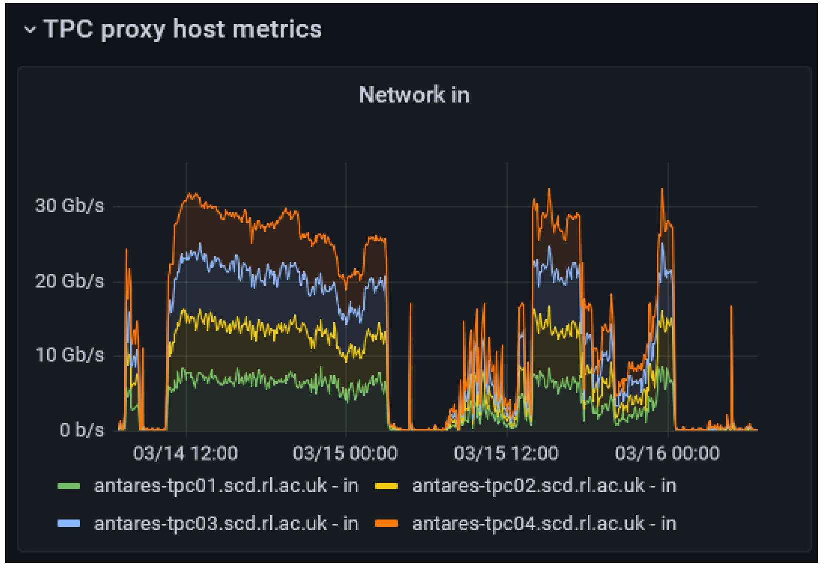
DDM monitoring of transfers to Echo

Data challenge lessons – TPC proxies

- Requirement for these was not well understood going into the first DC
  - CERN CTA do not need these as most TPC transfers are from trusted hosts (CERN EOS)
- Started with one repurposed EOS node as a proxy, added another by the end of the challenge week to cope with demand
- Going into production, we have four dedicated TPC nodes



Data challenge 1 setup



Production setup

Science and Technology Facilities Council

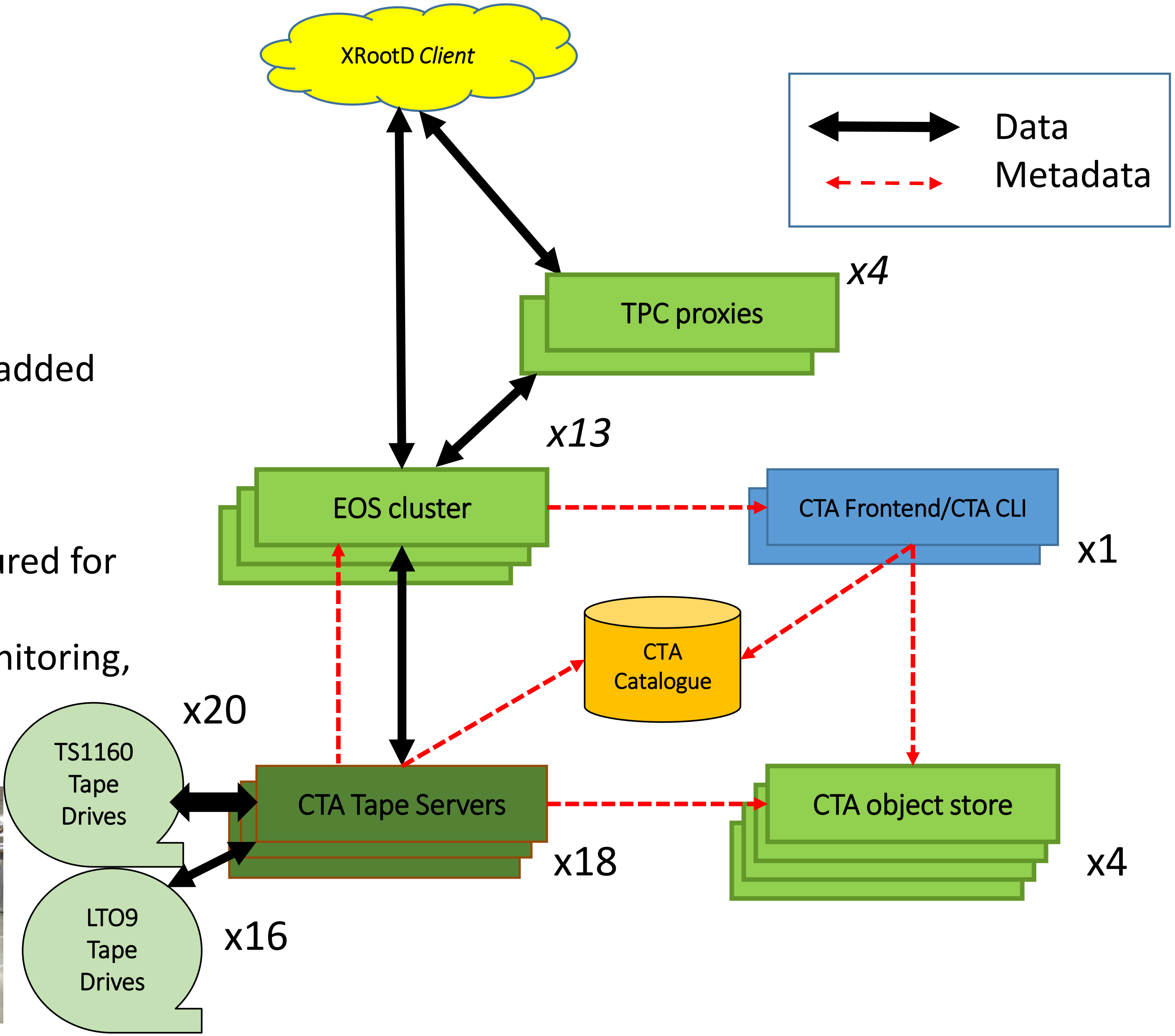
Antares production setup

Hardware changes:

- Dedicated TPC proxies configured (100Gb/s combined throughput available)
- Full compliment of EOS nodes provisioned
- Another 5 tape servers with TS1160 drives added
- 7 tape servers with LTO9 tape drives added

Software changes:

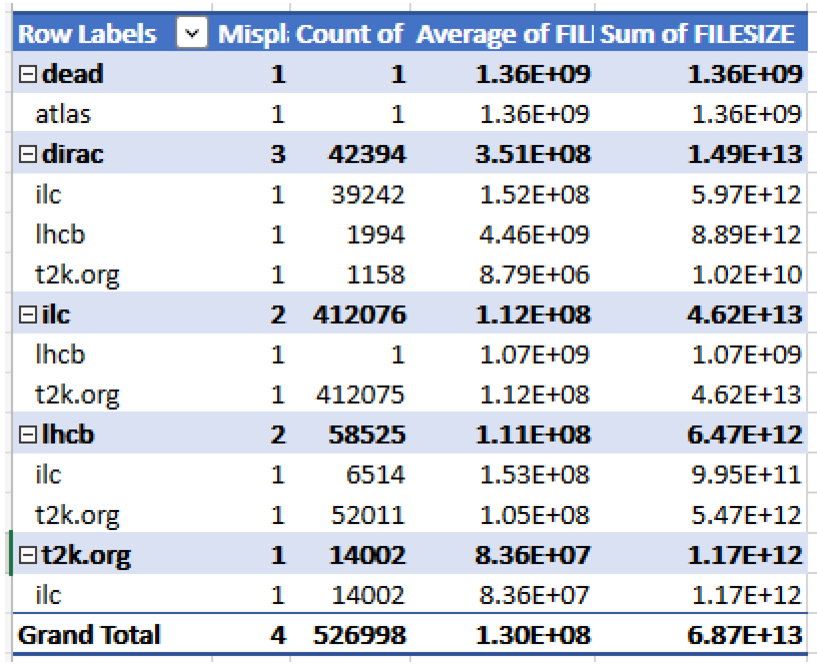
- Alice authentication configured and tested
- WebDav support (incl. TPC support) configured for LHCb needs
- Lots of production readiness changes – monitoring, alerting, DR



Science and Technology Facilities Council

Tier-1 Castor to Antares Migration

- Migration method:
  - ➢ Castor namespace injection to EOS
  - ➢ Castor tape metadata migration to CTA DB
- Migration pre-requisites:
  - ✓ Upgrade to CASTOR to 2.1.19-3
  - ✓ Import CASTOR DB schemas (NS,VMGR,STAGER) snapshot to the CTA DB
  - ✓ Review/modify PL/SQL scripts to be run on the schemas
  - ✓ Further namespace clean up (repack files to the right tape pools)
  - ✓ Set up a migration node to run the migration client tools
  - ✓ Estimate timings to be scheduled in the intervention plan

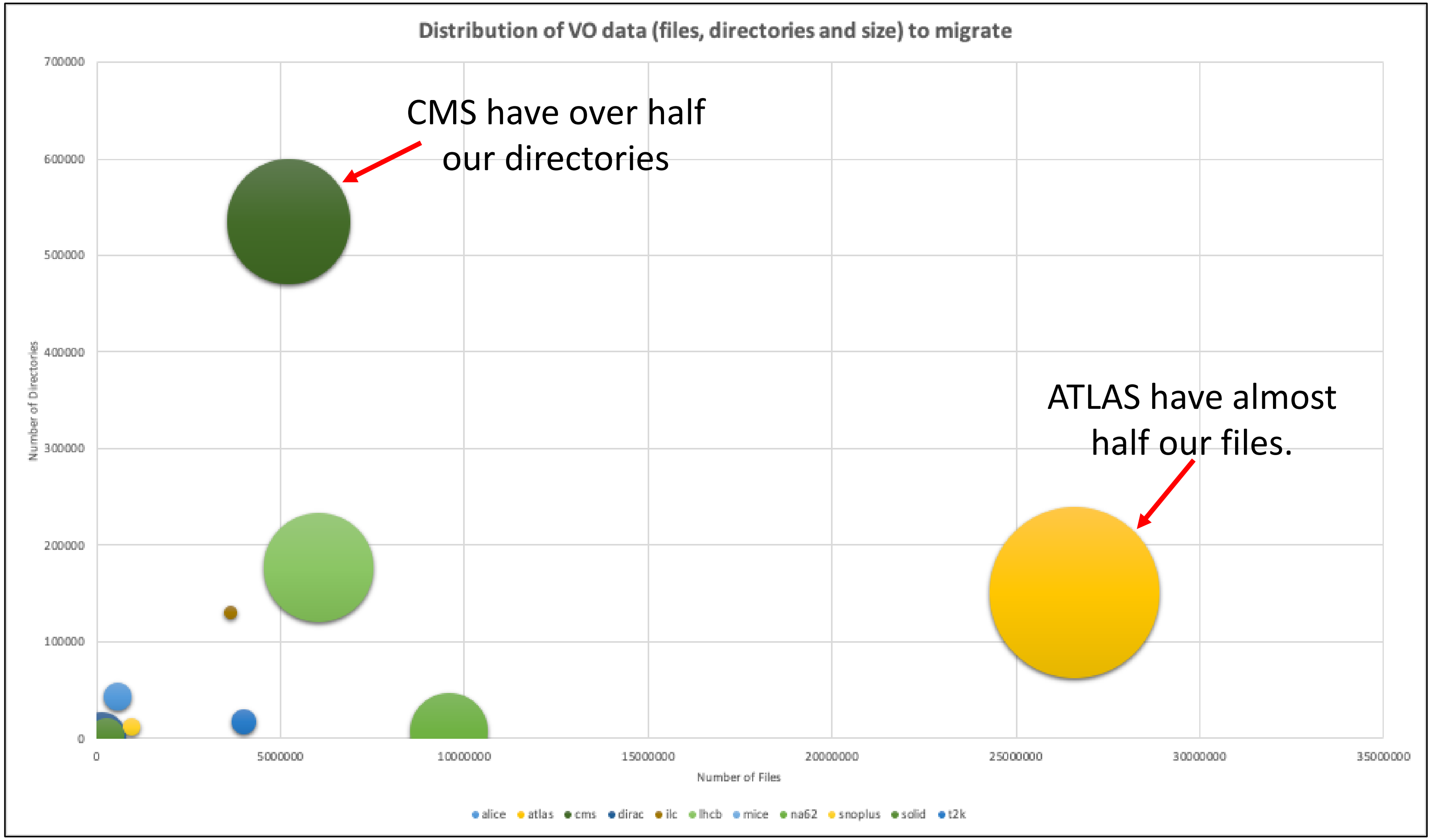


Row Labels	Mispl	Count of	Average of FILI	Sum of FILESIZE
⊟ **dead**	**1**	**1**	**1.36E+09**	**1.36E+09**
atlas	1	1	1.36E+09	1.36E+09
⊟ **dirac**	**3**	**42394**	**3.51E+08**	**1.49E+13**
ilc	1	39242	1.52E+08	5.97E+12
lhcb	1	1994	4.46E+09	8.89E+12
t2k.org	1	1158	8.79E+06	1.02E+10
⊟ **ilc**	**2**	**412076**	**1.12E+08**	**4.62E+13**
lhcb	1	1	1.07E+09	1.07E+09
t2k.org	1	412075	1.12E+08	4.62E+13
⊟ **lhcb**	**2**	**58525**	**1.11E+08**	**6.47E+12**
ilc	1	6514	1.53E+08	9.95E+11
t2k.org	1	52011	1.05E+08	5.47E+12
⊟ **t2k.org**	**1**	**14002**	**8.36E+07**	**1.17E+12**
ilc	1	14002	8.36E+07	1.17E+12
Grand Total	**4**	**526998**	**1.30E+08**	**6.87E+13**

Tier-1 CASTOR to Antares migration

Total:

- 1,079,217 dirs
- 57,011,928 files
- 70.5PB



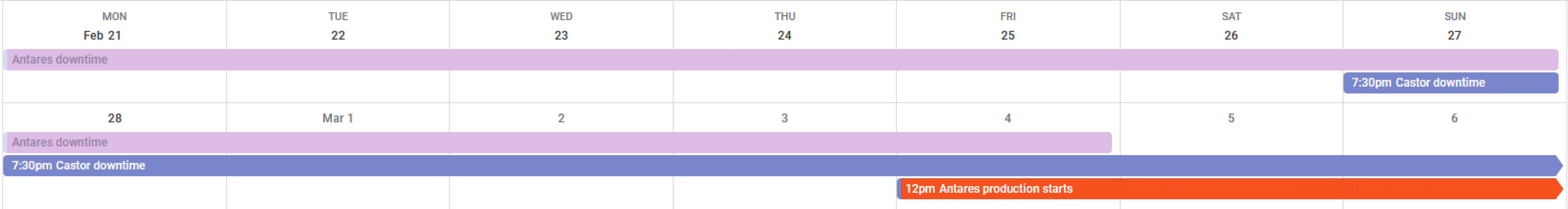
Distribution of VO data (files, directories and size) to migrate

CMS have over half our directories

ATLAS have almost half our files.

Castor to Antares Migration

- Actual migration – downtime Sunday pm to allow Castor to drain - to midday Thursday
- Backups of Castor taken and file lists created for VOs pre-migration – rollback checkpoint
- Two team members migrating all the VOs one at a time. File lists in EOS produced for each VO to compare with the Castor file list --> small numbers of anomalies recorded and investigated!
- Had to apply dir extended attributes (ACLs and CTA workflows) on the <u>*whole*</u> dir structure after migration
- Scripted applying across the whole namespace – ATLAS: 150,000 dirs, CMS: 535,000 dirs, LHCb: 177,000 dirs – and found that 70,000 directories was the maximum namespace size to apply the attributes without hitting the timeout limit
  - Required in a ~24 hours extension of the downtime
- Production traffic for some LHC VOs started on Friday afternoon

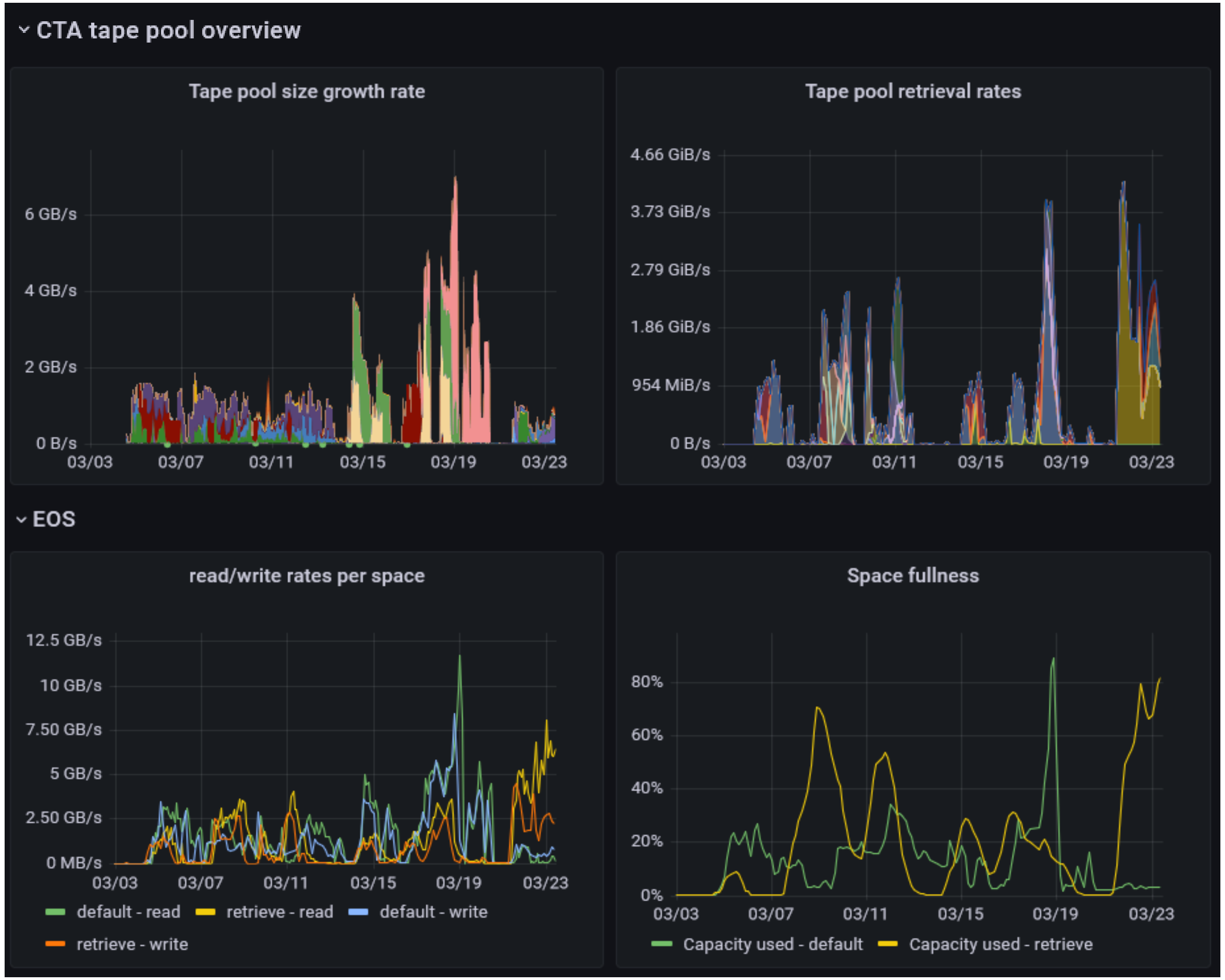


MON	TUE	WED	THU	FRI	SAT	SUN
Feb 21	22	23	24	25	26	27
Antares downtime						
						7:30pm Castor downtime
28	Mar 1	2	3	4	5	6
Antares downtime						
7:30pm Castor downtime						
				12pm Antares production starts		

Antares

Antares in production

- The first 19 days in production have been dominated by the data challenge
- Otherwise, things have been fairly smooth, but busy!
- A few issues to discuss…



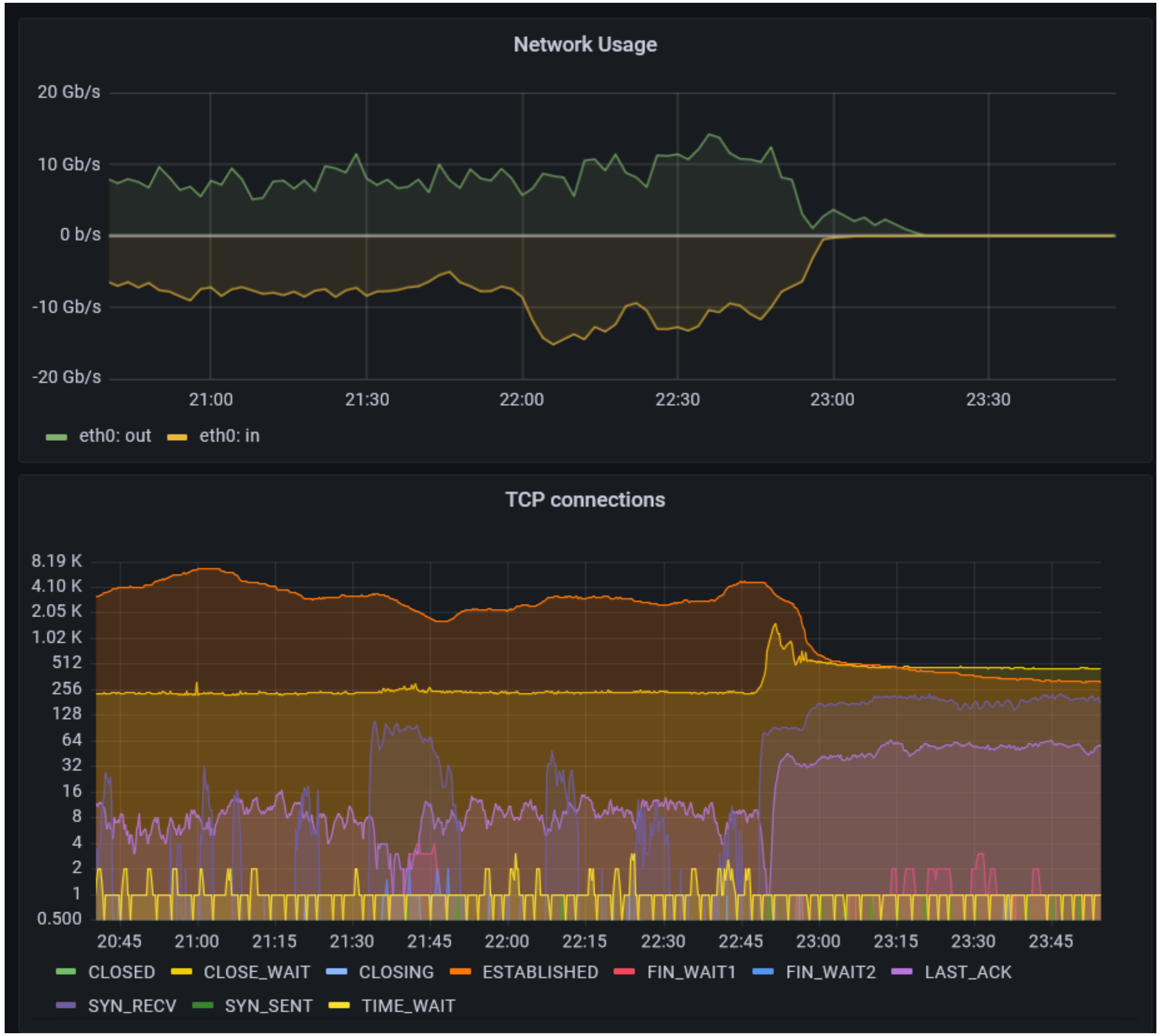
Science and Technology Facilities Council

Issues in production – Low LTO9 tape drive rates

- We found that the write performance of the LTO9 drives appeared to be approximately half of the expected rate
- This was tracked down to very poor network performance to the LTO9 tape servers
  - thought to be the ‘unusual’ network cables that had been used (Dell rather than Mellanox), but switching cables did not change rates
  - Network (ring buffer size) tuning vastly improved network rates
- Now we have expected network rates to tape servers, testing is ongoing to determine if LTO9 drive performance is similarly improved

Issues in production – EOS MGM stalls

- Occasional stalls of MGM node have been observed
- XRootD process still running happily, but all incoming connections fail
- Network config/tuning under scrutiny
  - Excessive packet discards seen in some cases



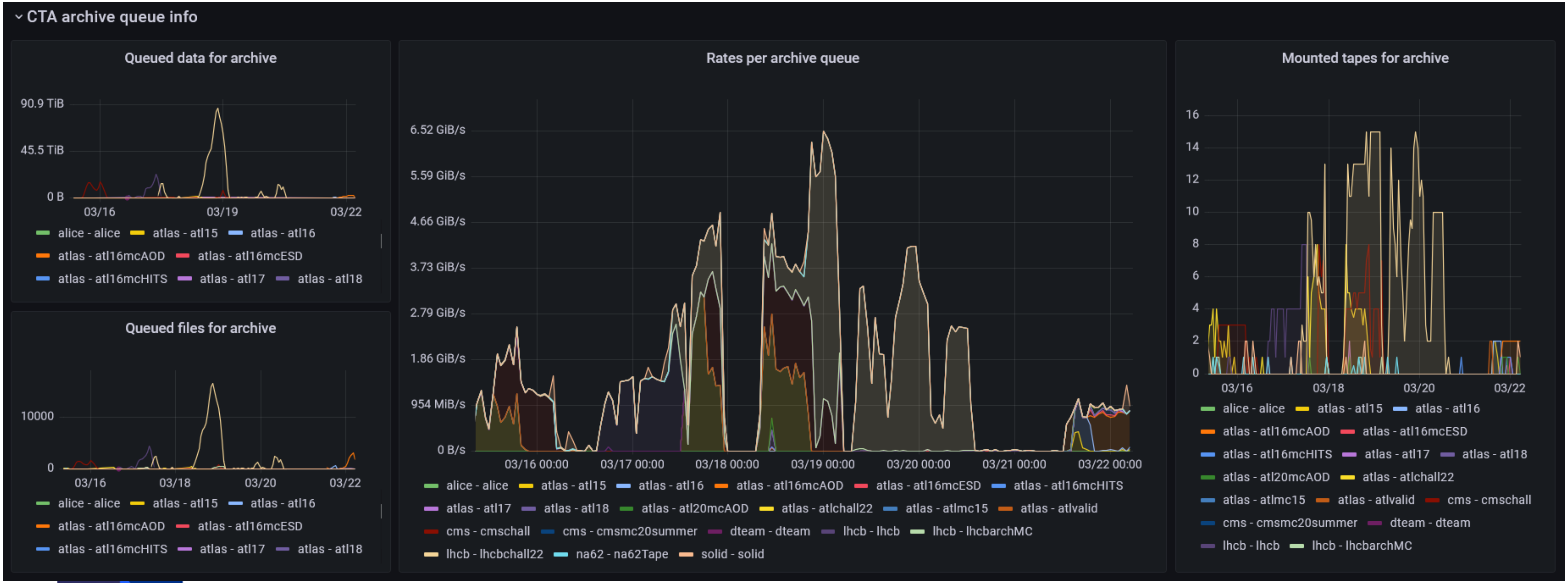
2022 LHC data challenge – operational perspective

Antares

- The presentation following this will cover how the current data challenge went from a VO perspective:
  - https://indico.cern.ch/event/1128343/contributions/4787155/

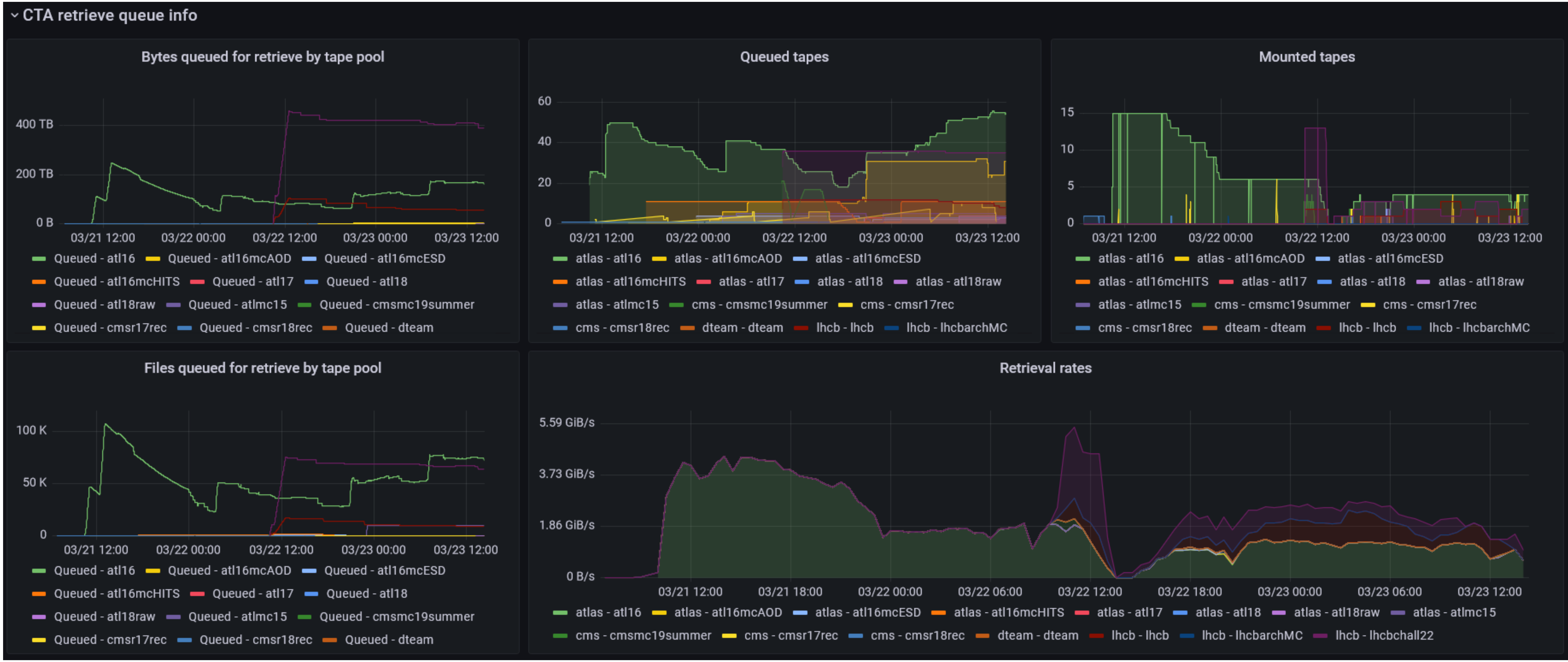
- I'd like to present a few thoughts on how Antares handled the challenge from our operational perspective

2022 LHC data challenge - archival



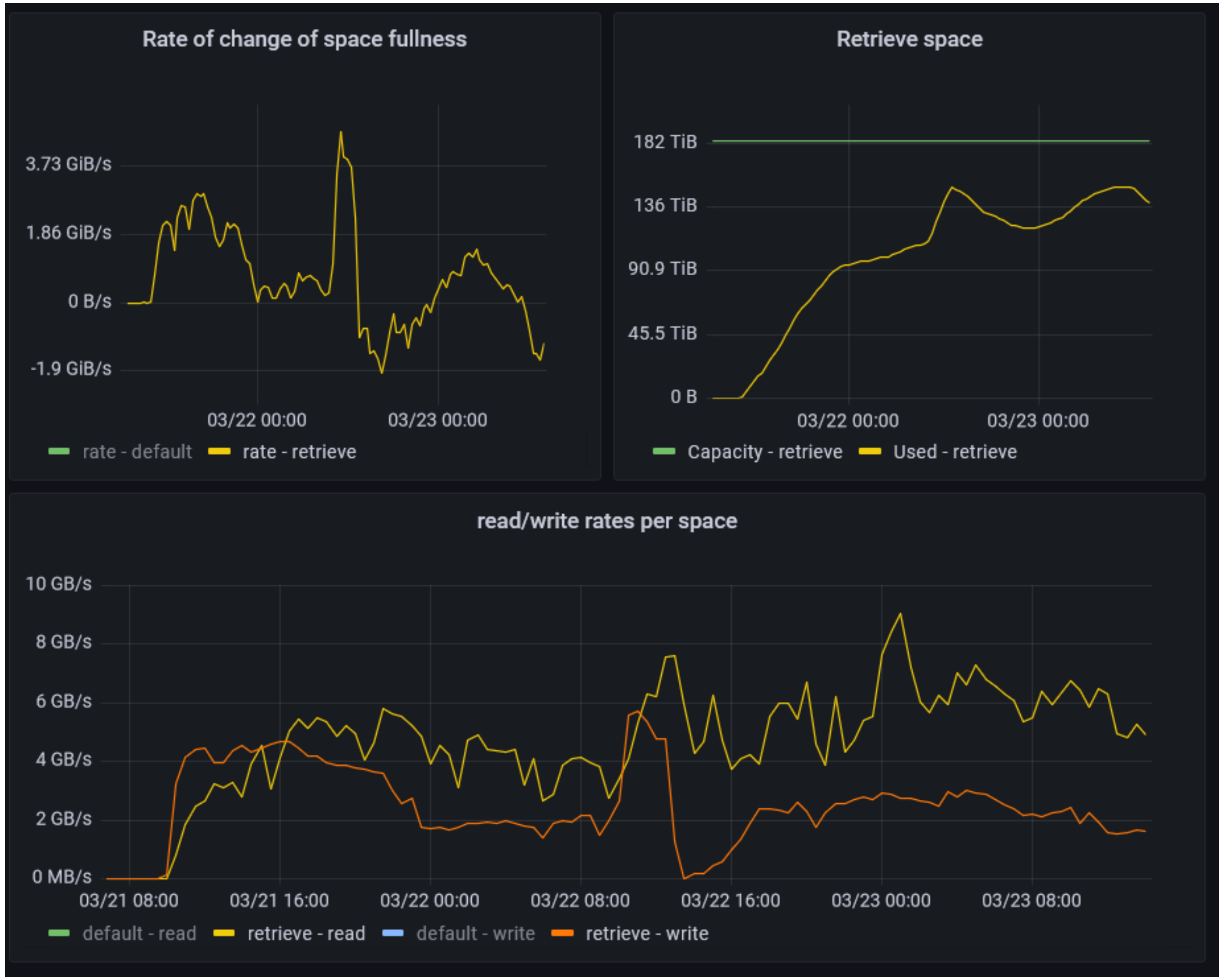
Science and Technology Facilities Council

2022 LHC data challenge – retrieval



2022 LHC data challenge – managing retrieval EOS space

- Archival has been proven to work well with the small, fast buffer model
- Retrieve relies on external systems to monitor status, copy out and clean up files as they come online
    - Any misconfiguration can result in things going wrong very quickly
    - Balancing retrieval pressures between different VOs with different access methods will be an interesting challenge



Science and
Technology
Facilities Council

Next steps

- Ensure access for non LHC VO's is working as expected (ongoing)
- Upgrade EOS/CTA to the versions deployed at CERN
- Upgrade to EOS5
- Prepare and execute the migration of CASTOR Facilities
- Migrate the CTA Catalogue from Oracle to PostgreSQL

Antares

Thanks

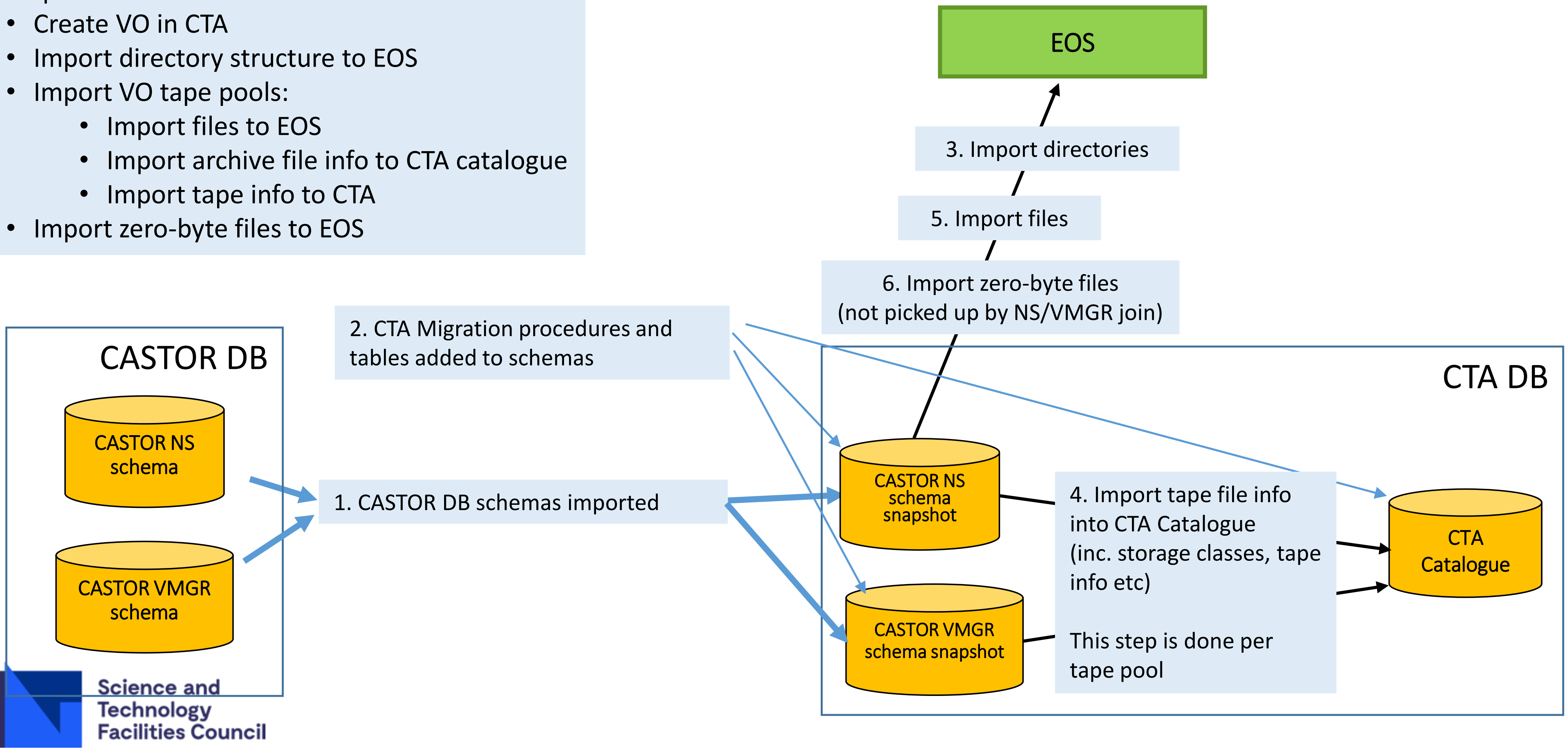
- Questions?

Science and
Technology
Facilities Council

Tier-1 Castor to Antares Migration

Steps needed for each VO:

- Create VO in CTA
- Import directory structure to EOS
- Import VO tape pools:
  - Import files to EOS
  - Import archive file info to CTA catalogue
  - Import tape info to CTA
- Import zero-byte files to EOS



CASTOR DB

CASTOR NS schema

CASTOR VMGR schema

1. CASTOR DB schemas imported

2. CTA Migration procedures and tables added to schemas

EOS

3. Import directories

5. Import files

6. Import zero-byte files (not picked up by NS/VMGR join)

CTA DB

CASTOR NS schema snapshot

CASTOR VMGR schema snapshot

4. Import tape file info into CTA Catalogue (inc. storage classes, tape info etc)

This step is done per tape pool

CTA Catalogue