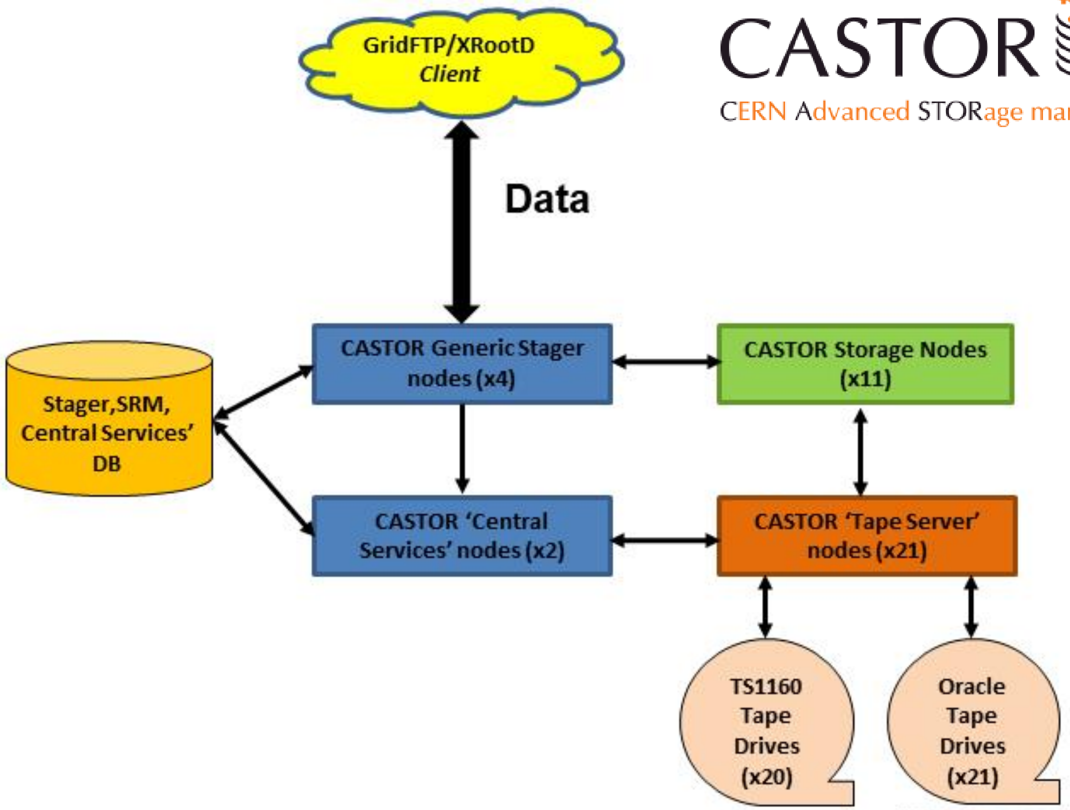
RAL CTA Deployment Update

Tom Byrne, George Patargias
2nd September 2021

Motivation

- CASTOR has provided the tape archive service at RAL since 2006
- Designed and maintained by CERN who have now migrated to CTA
- CTA has a number of benefits for RAL, including:
  - Opportunity to migrate with some or all data in place
  - Continue our relationship with the CASTOR/CTA team at CERN

CASTOR
CERN Advanced STORage manager



Science and
Technology
Facilities Council

Recent progress highlights

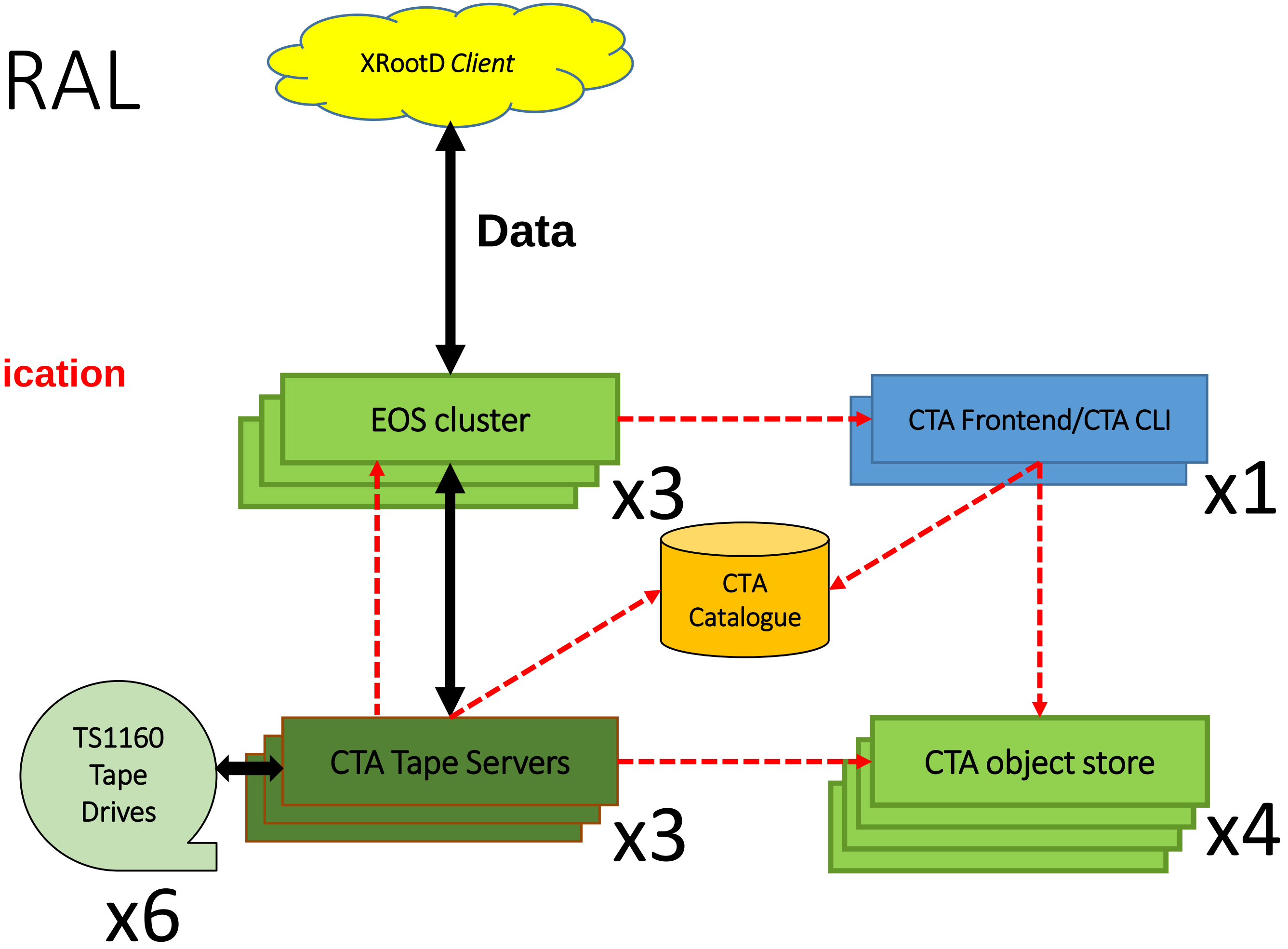
- Full EOS+CTA stack on production hardware at RAL
  - lots of work done, tearing up and tearing down instances
  - gearing up to run in production, better to learn from mistakes now!
- Lots of EOS benchmarking
- Hitting important milestones:
  - client to tape (and back again) demonstrated ☺
  - (internal) VO access demonstrated
- Preparation for VO data challenges

```
[root@cta-adm ~]# eos ns
# ------------------------------------------------------------------------
# Namespace Statistics
# ------------------------------------------------------------------------
ALL      Files                              3592 [booted] (0s)
ALL      Directories                        22
ALL      Total boot time                    0 s
# ------------------------------------------------------------------------
ALL      Replication                        is_master=true master_id=cta-
eos01.scd.rl.ac.uk:1094
# ------------------------------------------------------------------------
...
# ------------------------------------------------------------------------
ALL      tapeenabled                        true
ALL      tgc.stats=stagerrms                default=0 retrieve=0
ALL      tgc.stats=queuesize                default=3583 retrieve=1
ALL      tgc.stats=totalbytes               default=30711109189632 retrieve=
ALL      tgc.stats=availbytes               default=15351459610624 retrieve=
ALL      tgc.stats=qrytimestamp             default=1630007857 retrieve=1630
# ------------------------------------------------------------------------
[root@cta-adm ~]#
```

Science and Technology Facilities Council

EOS + CTA stack at RAL

as of August 2021



Tier-1 Spectra Logic Tfinity tape library

Two completely separate setups of this size deployed:

1. Internal deployment testing
2. VO and user testing

VO data challenges

- Upcoming VO driven data challenges to ensure all custodial data storage systems are ready for Run 3 rates
  - Good opertunity to validate CTA performance with these tests
- Regular discussions with VO liaisons to ensure we will meet requirements for the tests
- Planning to deploy a larger EOS+CTA stack for the testing. Important factors are tape drive data rate, EOS node capacity, and EOS node data rate
  - 400MB/s throughput per tape drive
  - 32TB capacity per EOS node
  - ~2GB/s throughput per EOS node

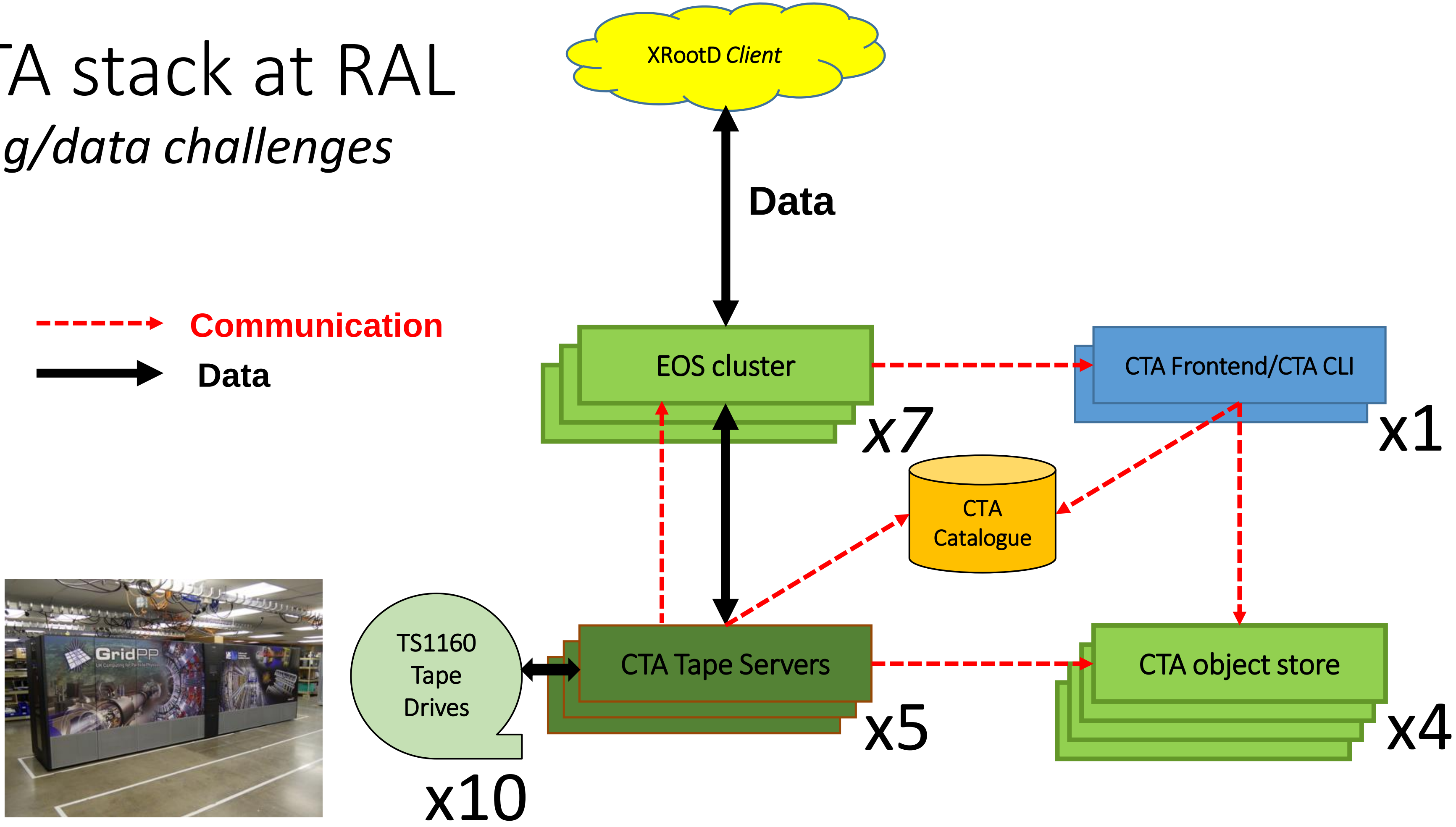
VO	Throughput (GB/s)
ALICE	0.08
ATLAS	1.6
CMS	0.9
LHCb	1.46
Total	4.04

required rates for data challenges

Science and Technology Facilities Council

EOS + CTA stack at RAL

for VO testing/data challenges



Tier-1 Spectra Logic Tfinity tape library

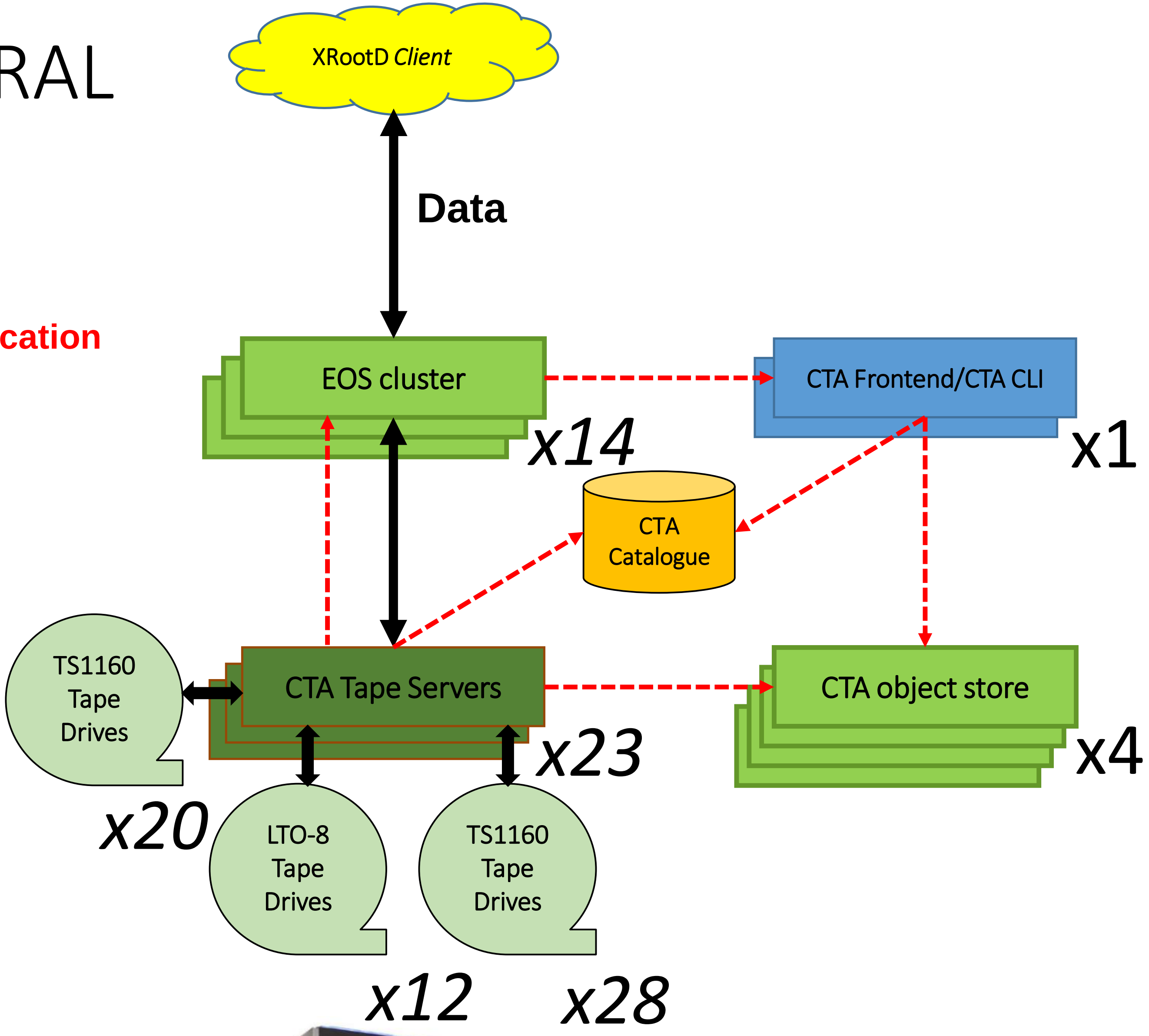
External testing setup expanded by 3 EOS nodes and 4 tape drives to meet buffer size and tape bandwidth requirements

Science and Technology Facilities Council

EOS + CTA stack at RAL

Production scale

Communication
Data



Tier-1 Spectra Logic Tfinity tape library

Facilities SpectraTfinity tape library

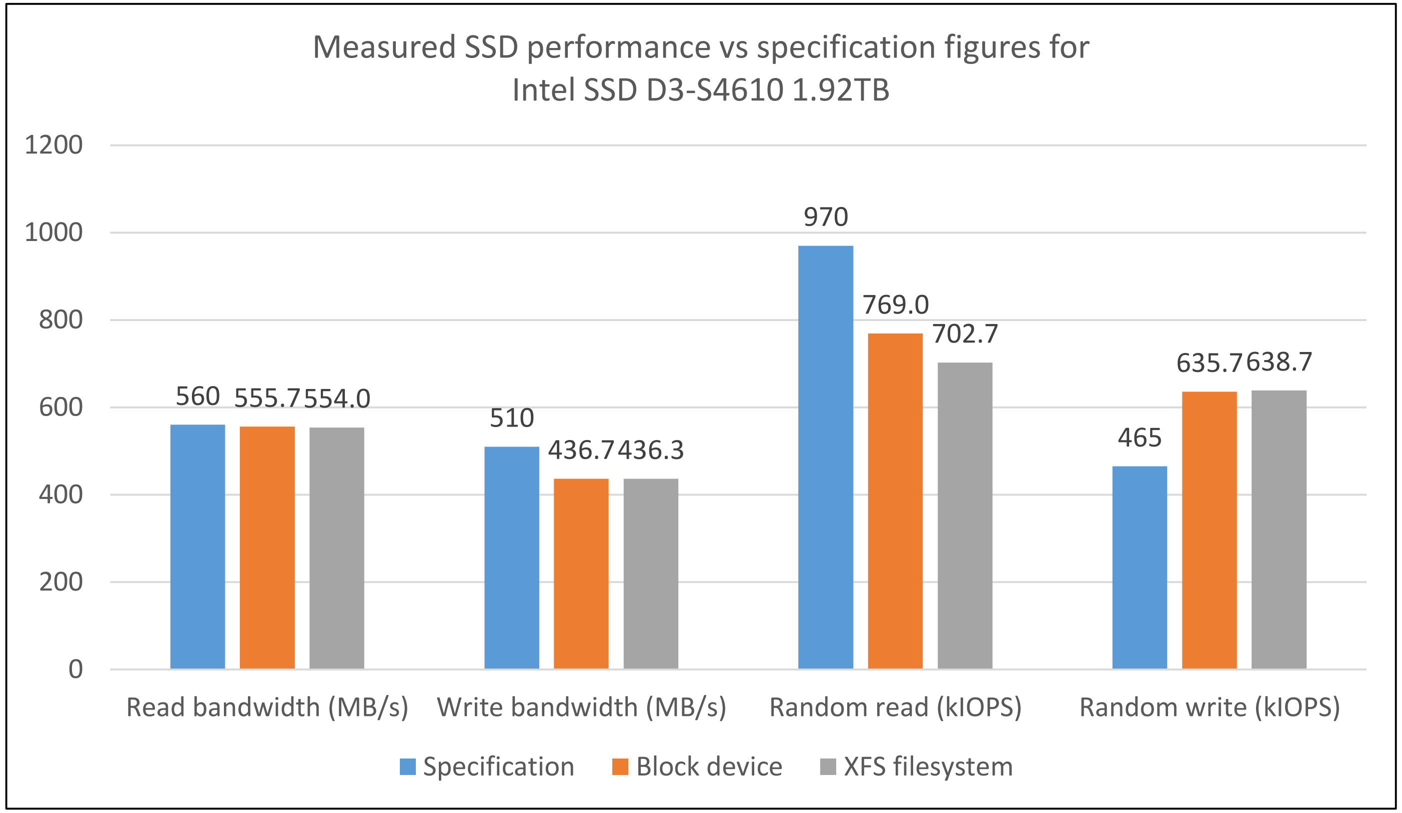
Science and Technology Facilities Council

EOS hardware benchmarking

- The CERN CTA model relies on a small fast disk buffer (EOS)
  - Individual node streaming performance needs to be good for the model to work well
- The hardware bought for EOS has appropriate hardware specs, but always good to validate performance!
- Start at the lowest level and work up:
  - disk
  - node
  - storage software layer (EOS)
- Understanding hardware performance is important to avoid optimization headaches down the line!

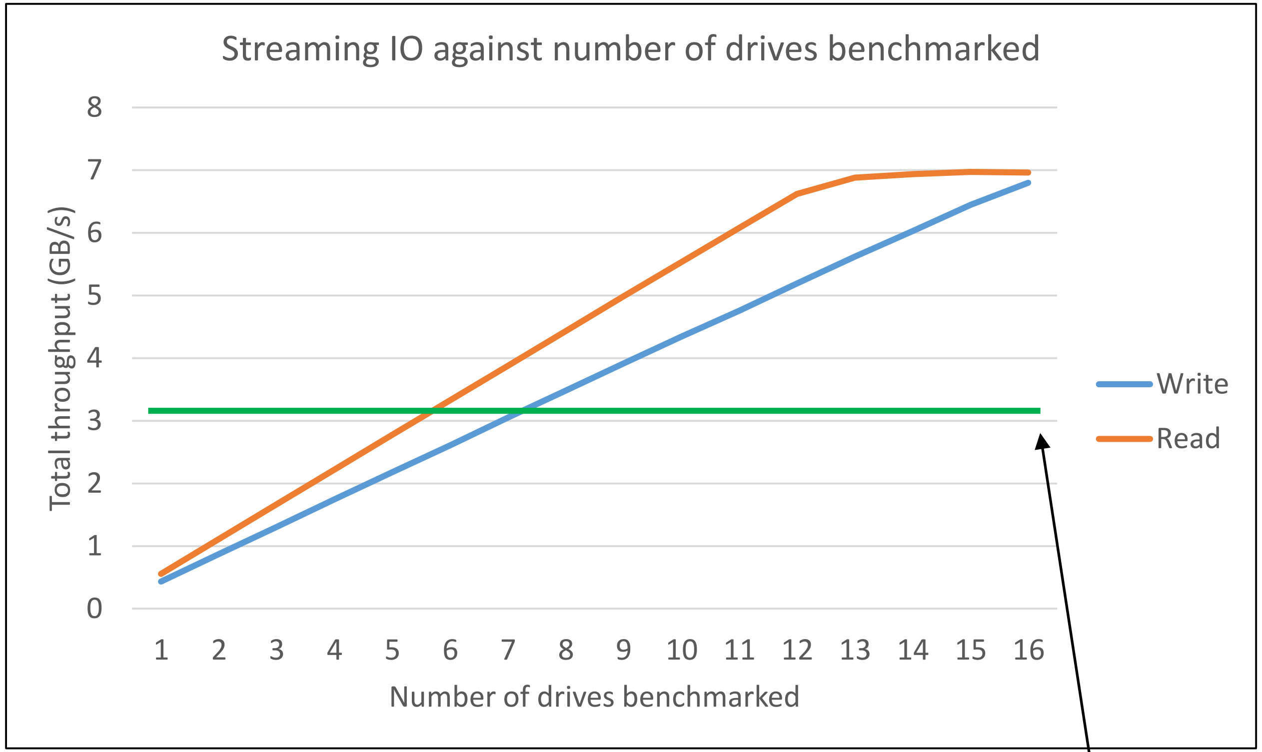
Single disk performance

- No serious surprises here
  - and no tuning necessary
- IOPS numbers were lower, and then higher than spec, which was interesting, but not hugely relevant for the CTA use case
  - bandwidth is important for the tape buffer
- All testing done with FIO and stock SL7 kernel



Single workload aggregate performance

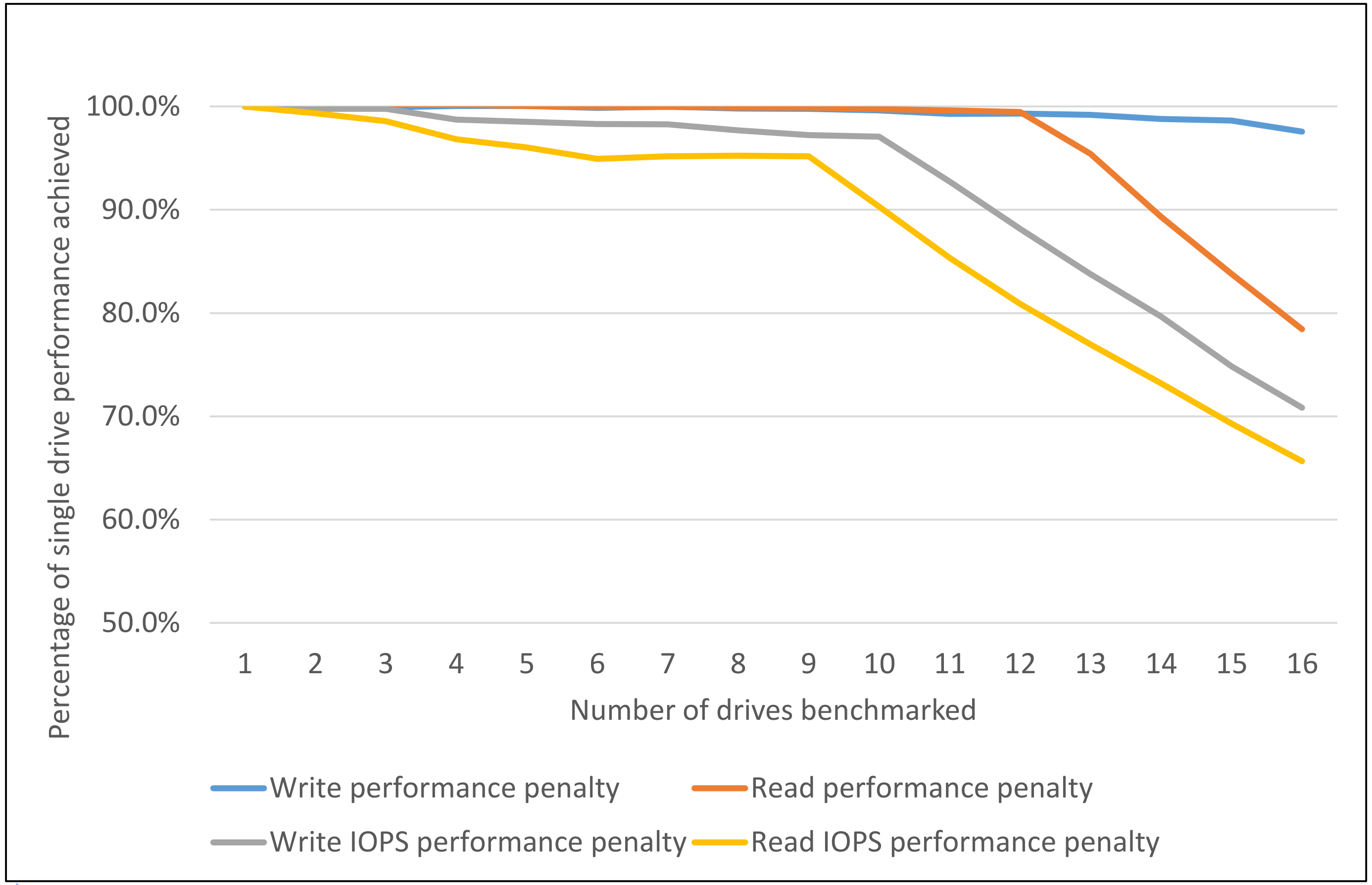
- Performance scaled well
- Aggregate throughput limited by PCIe 3.0 8x HBA
  - Limit significantly higher than network capabilities
  - Will be less of a problem with next gen hardware – but current limit is much higher then network limits, so not an issue here



Node network capability - 25Gb/s

Science and Technology Facilities Council

Aggregate performance

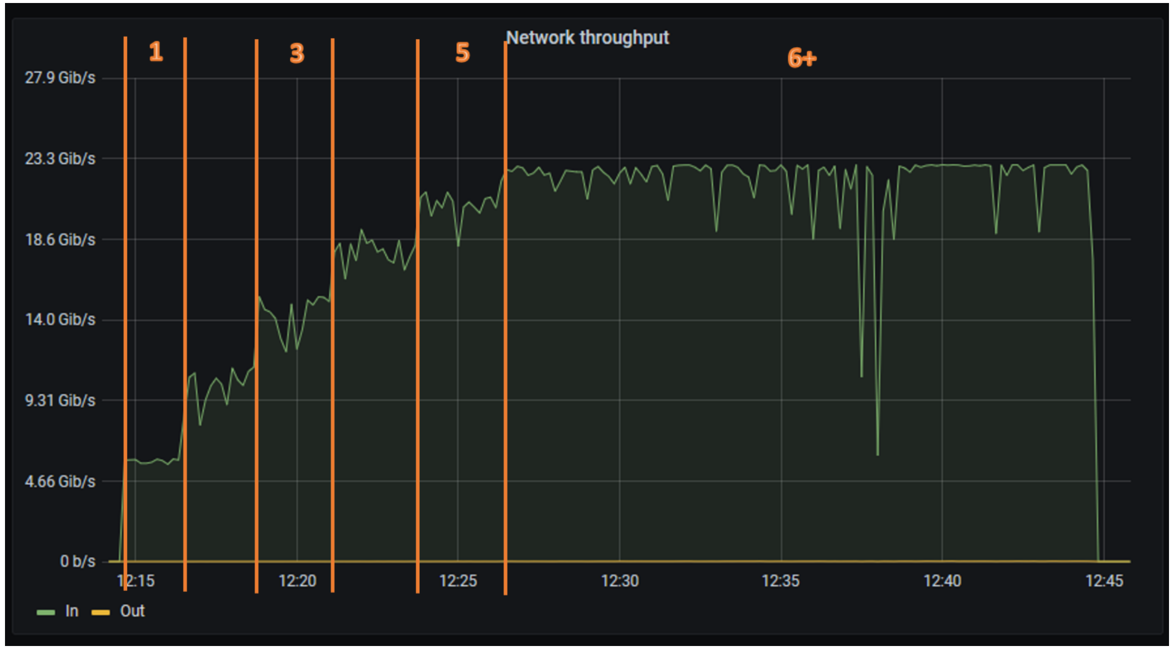


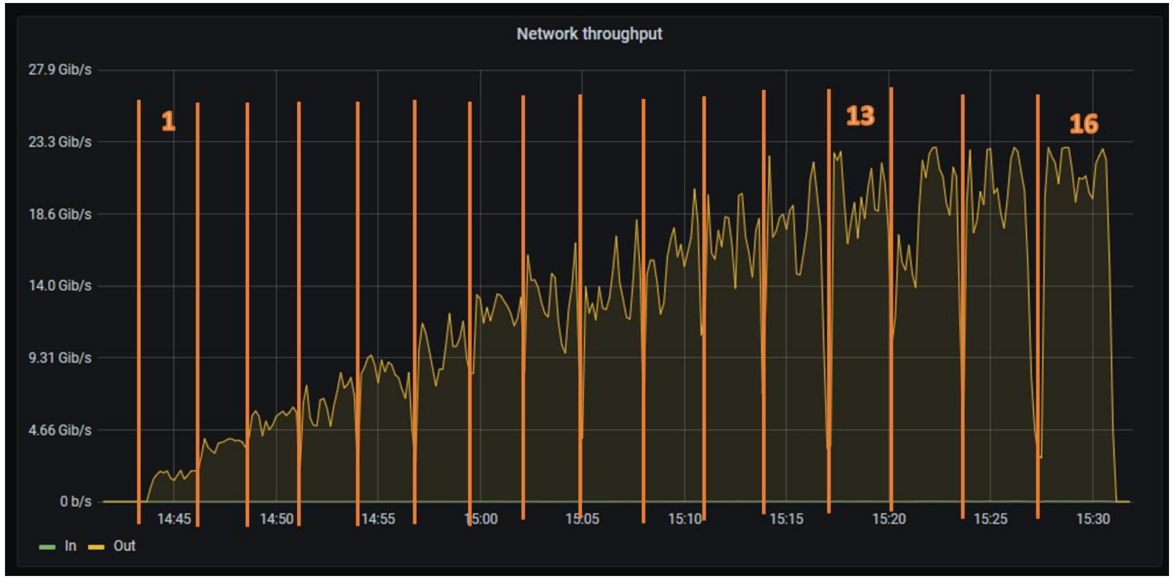
- The plot shows the percentage of single drive performance achieved when benchmarking multiple drives
- For streaming workloads, little aggregate penalty seen until the 7GB/s mark, very encouraging
- For IOPS intensive workload, significantly more contention seen, the sweet spot seemed to be ~8 drives per node.
  - Again, IOPS not a concern for the use-case, so little time spent investigating

Science and Technology Facilities Council

Single node EOS cluster benchmarking

- XRootD write performance matches FIO testing
    - 25Gb interface saturation with 6 XRootD transfers
- XRootD read performance lower than synthetic tests
    - Still capable of easily saturating network interface
- EOS <-> ECHO large scale throughput testing coming soon
- **EOS buffer performance exceeds requirements for EOS+CTA use case**





Science and Technology Facilities Council

Demonstrating Archive/Retrieve functionality

Archive

```
[vwa13372@lcgui05 ~]$ xrdcp ./2gb root://cta-eos14.scd.rl.ac.uk//eos/antares/dteam/tape/tom-test-2gb-3
[2GB/2GB][100%][==================================================][292.6MB/s]
[vwa13372@lcgui05 ~]$
```

```
[root@cta-eos14 ~]# eos info /eos/antares/dteam/tape/tom-test-2gb-3
File: '/eos/antares/dteam/tape/tom-test-2gb-3'  Flags: 0644
  Size: 2147483648
Modify: Fri Aug 27 10:39:16 2021 Timestamp: 1630057156.554637000
CUid: 1000 CGid: 1000 Fxid: 00000066 Fid: 102 Pid: 20 Pxid: 00000014
XStype: adler    XS: 1f 32 c1 d4    ETAGs: "27380416512:1f32c1d4"
Layout: replica Stripes: 1 Blocksize: 4k LayoutId: 00100012 Redundancy: d1::t0
  #Rep: 1
```

File on disk

no.	fs-id	host	schedgroup	path	boot	configstatus	drain	active	geotag
0	7	cta-eos14.scd.rl.ac.uk	default.0	/eos/data07	booted	rw	nodrain	online	rack1

```
[root@cta-eos14 ~]#
```

```
[root@cta-front02 ~]# cta-admin showqueues
type   tapepool    vo logical library vid files queued data queued oldest age priority min age read max drives write max drives cur. mounts cur. files cu
ArchiveForUser dteam_test dteam                - -           1     2.1G         1        1     5              2               2          1
[root@cta-front02 ~]#
```

Archiving

```
[root@cta-eos14 ~]# eos info /eos/antares/dteam/tape/tom-test-2gb-3
File: '/eos/antares/dteam/tape/tom-test-2gb-3'  Flags: 0644
  Size: 2147483648
Modify: Fri Aug 27 10:39:16 2021 Timestamp: 1630057156.554637000
CUid: 1000 CGid: 1000 Fxid: 00000066 Fid: 102 Pid: 20 Pxid: 00000014
XStype: adler    XS: 1f 32 c1 d4    ETAGs: "27380416512:1f32c1d4"
Layout: replica Stripes: 1 Blocksize: 4k LayoutId: 00100012 Redundancy: d0::t1
  #Rep: 1
TapeID: 4294967324 StorageClass: dteam_test
*******
[root@cta-eos14 ~]
```

File on tape!

Retrieve

```
[vwa13372@lcgui05 ~]$ xrdfs root://cta-eos14.scd.rl.ac.uk prepare -s /eos/antares/dteam/tape/tom-test-2gb-3
0446c0a89255:31ccfe1d.6127afeb:4:1630057605
[vwa13372@lcgui05 ~]$ xrdfs root://cta-eos14.scd.rl.ac.uk query prepare 0446c0a89255:31ccfe1d.6127afeb:4:1630057605 /eos/antares/dteam/tape/tom-test-2gb-3
{"request_id":" 0446c0a89255:31ccfe1d.6127afeb:4:1630057605 ","responses":[{"path":"/eos/antares/dteam/tape/tom-test-2gb-3","exists":true,
"path_exists":true,"on_tape":true,"online":false,"requested":true,"has_reqid":true,"req_time":"1630057605","error_text":""}]}
[vwa13372@lcgui05 ~]$
```

File not ready

```
[root@cta-front02 ~]# cta-admin showqueues
type     tapepool    vo logical library    vid files queued data queued oldest age priority min age read max drives write max drives cur. mounts cur. files c
Retrieve dteam_test dteam      asterix_ts CT4852            0            0          0        0       0                2                 2           1
[root@cta-front02 ~]#
```

```
[root@cta-eos14 ~]# eos info /eos/antares/dteam/tape/tom-test-2gb-3
[snip]
Layout: replica Stripes: 1 Blocksize: 4k LayoutId: 00100012 Redundancy: d1::t1
  #Rep: 2
TapeID: 4294967324 StorageClass: dteam_test
```

Back on disk

no.	fs-id	host	schedgroup	path	boot	configstatus	drain	active	geotag
1	42	cta-eos99.scd.rl.ac.uk	**retrieve.0**	/eos/data12	booted	rw	nodrain	online	rack1

```
[root@cta-eos14 ~]#
```

```
[vwa13372@lcgui05 ~]$ xrdfs root://cta-eos14.scd.rl.ac.uk query prepare 0446c0a89255:31ccfe1d.6127afeb:4:1630057605
/eos/antares/dteam/tape/tom-test-2gb-3 | jq '.responses[0].online'
true
[vwa13372@lcgui05 ~]$ xrdcp root://cta-eos14.scd.rl.ac.uk//eos/antares/dteam/tape/tom-test-2gb-3 - > /dev/null
[2GB/2GB][100%][==================================================][1024MB/s]
[vwa13372@lcgui05 ~]$
```

File ready for retrieval

Science and Technology Facilities Council

Bonus progress – what’s in a name?

- Calling CTA @ RAL “CTA” is confusing
  - Having a ‘unique’ name for the tape archival service at RAL is good
- Many focus groups and sleepless nights later:

Antares

- ***A N***ew ***T***ape ***AR***chiv***E*** (for ***S***TFC)
  - Fits with the data services groups ‘name of stars’ naming scheme
  - Thanks to Matt Heath in our group for the name and backronym…
  - and to Helen Towrie in CLF for the logo

Short term plans

- Lots of effort going into understanding VO access requirements
  - RAL CASTOR has a lot of the answers, but keen to avoid unnecessary legacy as much as possible
- Basic VO testing has started
  - Mainly ensuring our authn/z is working for VOs for the upcoming challenges and other testing
  - CTA -> Echo (dev) TPC tested and working
- External access for Antares is imminent, paving the way to full external VO testing
  - This is coupled with RAL Tier-1 networking changes
- Migration planning and testing ongoing

Migration plan

Hardware networked ✓
Q1/2021

EOSCTA installed ✓
Q1/2021

CASTOR upgrade ✓
Q2/2021

Functional testing ✓
Q2/2021

Migration to Spectra completed ✓
Q2/2021

VO testing on CTA
Q2-Q3/2021

VO migration
Q3/2021-Q1/2022

Science and Technology Facilities Council

Migration details

CASTOR
CERN Advanced STORage manager

- Merging two CASTORs into one CTA means changing fileIDs of one instance to avoid collisions
  - Changing fileIDs means rewriting tapes ☹
- Current plan to migrate the WLCG CASTOR to CTA, leaving fileIDs intact
- The Facilities Castor data can be 'migrated' onto the CTA at a later date
- Lots of dry runs and testing!

"Castor Canadensis preparing for migration"
Photo by Steve from washington, dc, usa - American Beaver, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=3963858

UKRI Science and Technology Facilities Council

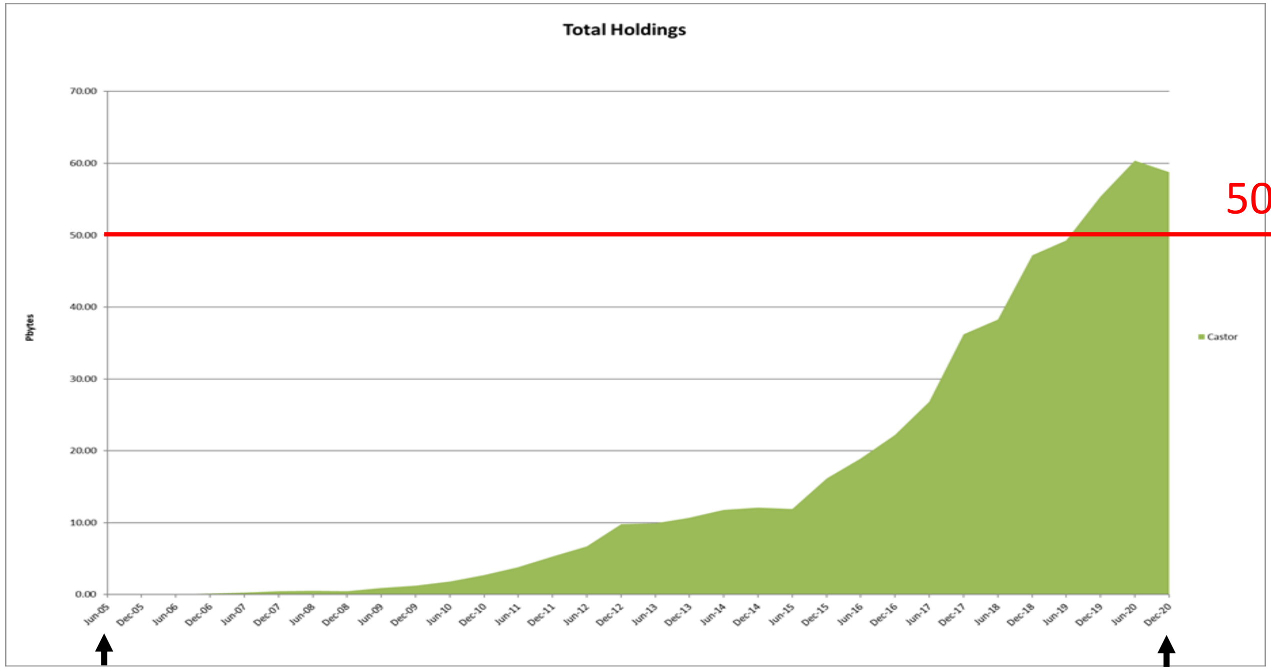
Thanks

- Questions?

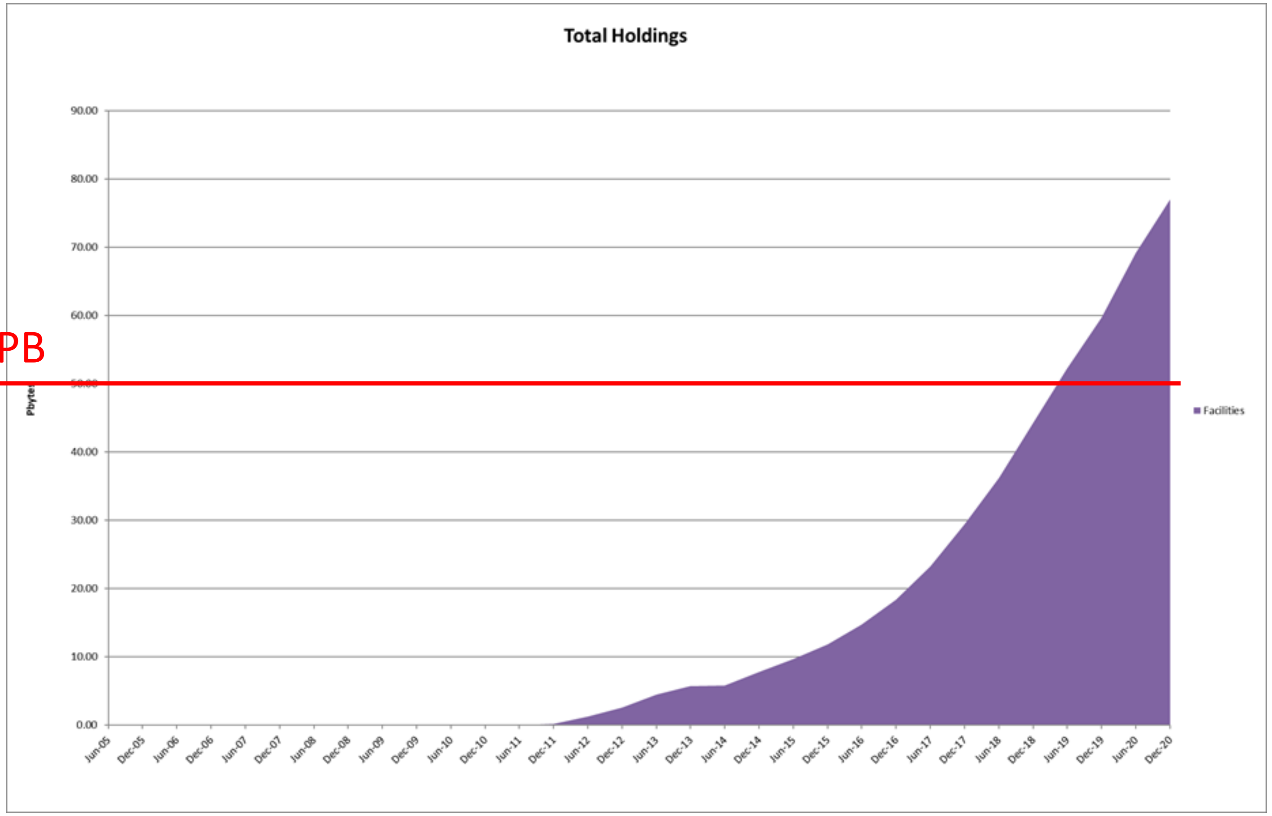
Science and
Technology
Facilities Council

Backups

STFC Tier-1/Facilities Castor Data Volumes



2005 2021

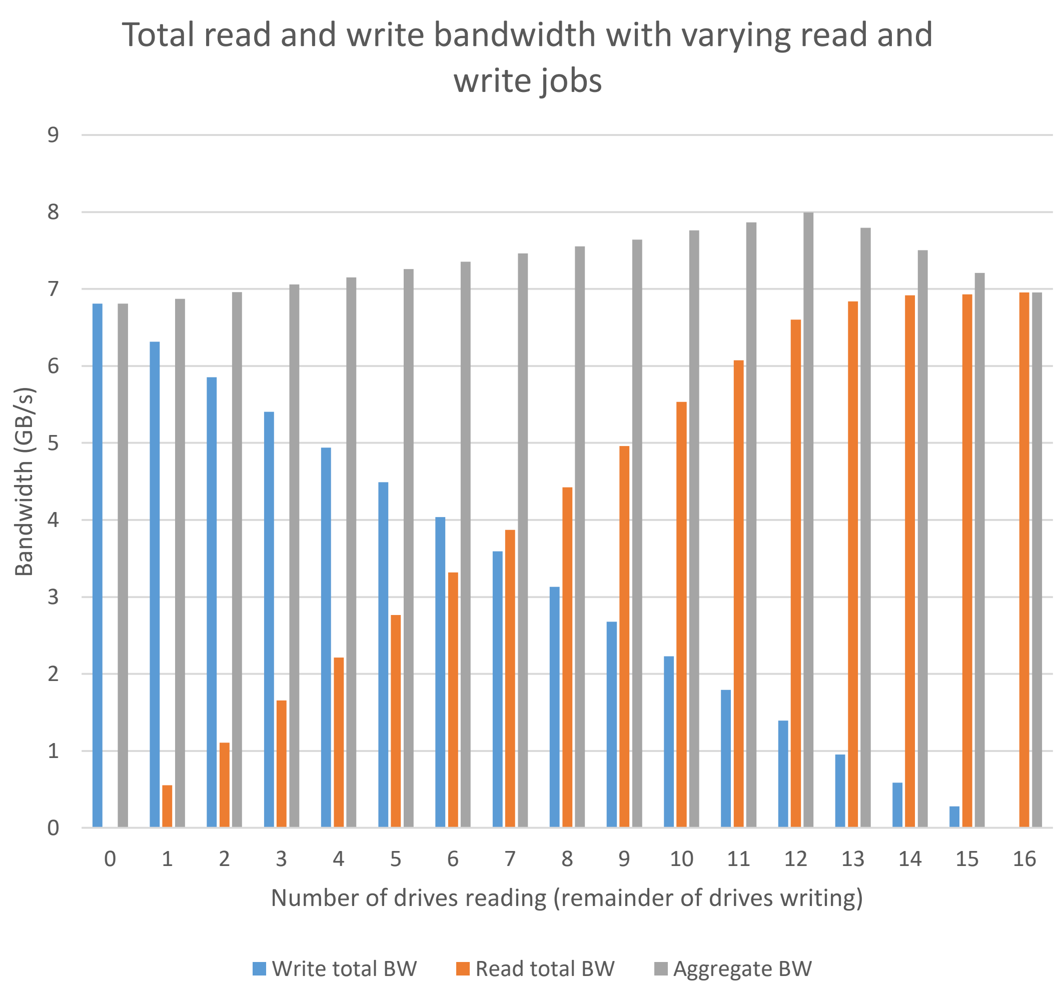


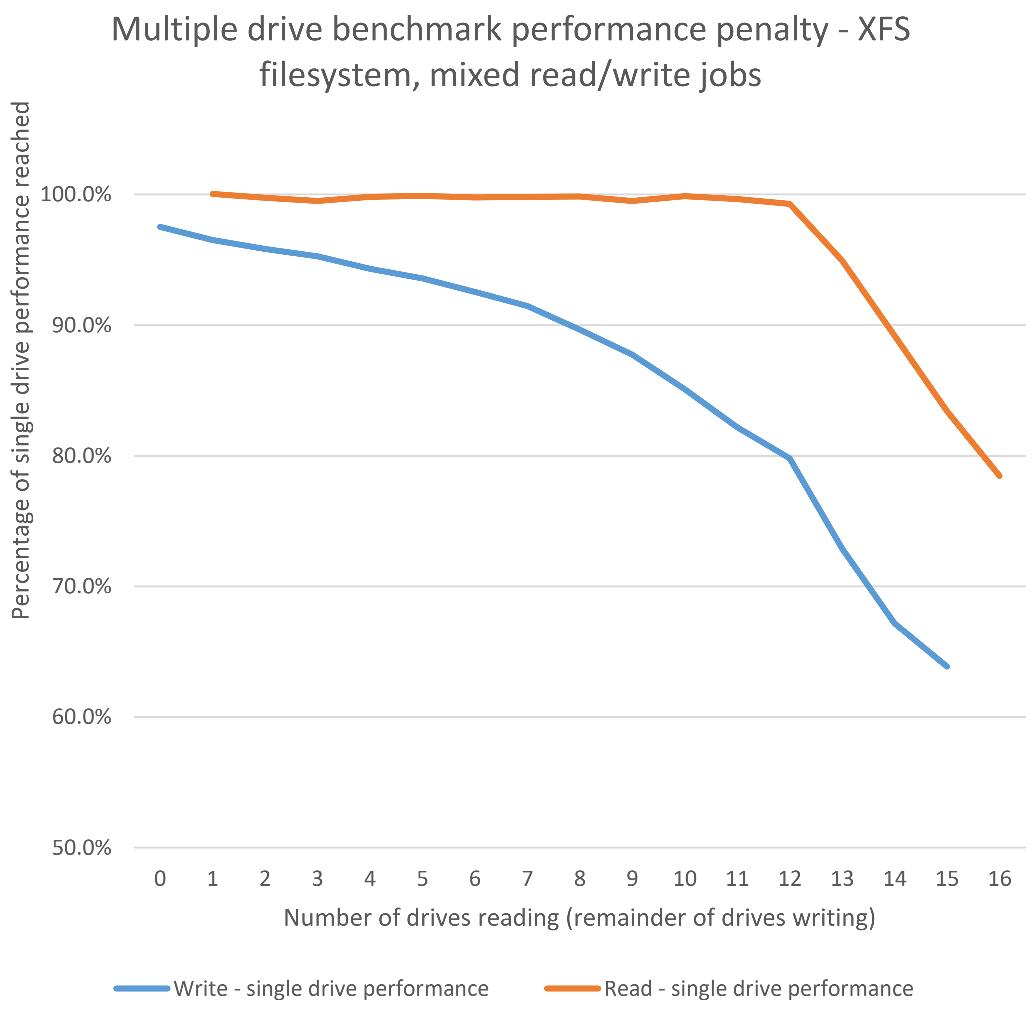
- **Tier-1/Facilities tape holdings > 130PB**
- **Growth rates: 0.8PB/month (Tier-1), 1.1PB/month (Facilities)**

Science and Technology Facilities Council

```
[root@cta-adm ~]# eos ns
# ------------------------------------------------------------------------------------
# Namespace Statistics
# ------------------------------------------------------------------------------------
ALL      Files                            3592 [booted] (0s)
ALL      Directories                      22
ALL      Total boot time                  0 s
# ------------------------------------------------------------------------------------
ALL      Replication                      is_master=true master_id=cta-
eos01.scd.rl.ac.uk:1094
# ------------------------------------------------------------------------------------
...
# ------------------------------------------------------------------------------------
ALL      tapeenabled                      true
ALL      tgc.stats=stagerrms              default=0 retrieve=0
ALL      tgc.stats=queuesize              default=3583 retrieve=1
ALL      tgc.stats=totalbytes             default=30711109189632 retrieve=61422218379264
ALL      tgc.stats=availbytes             default=15351459610624 retrieve=61416772272128
ALL      tgc.stats=qrytimestamp           default=1630007857 retrieve=1630007857
# ------------------------------------------------------------------------------------
[root@cta-adm ~]#
```

Science and
Technology
Facilities Council





Science and Technology Facilities Council