Science and
Technology
Facilities Council

CTA Deployment at RAL

George Patargias, Tom Byrne and Alison Packer

9th February, 2021

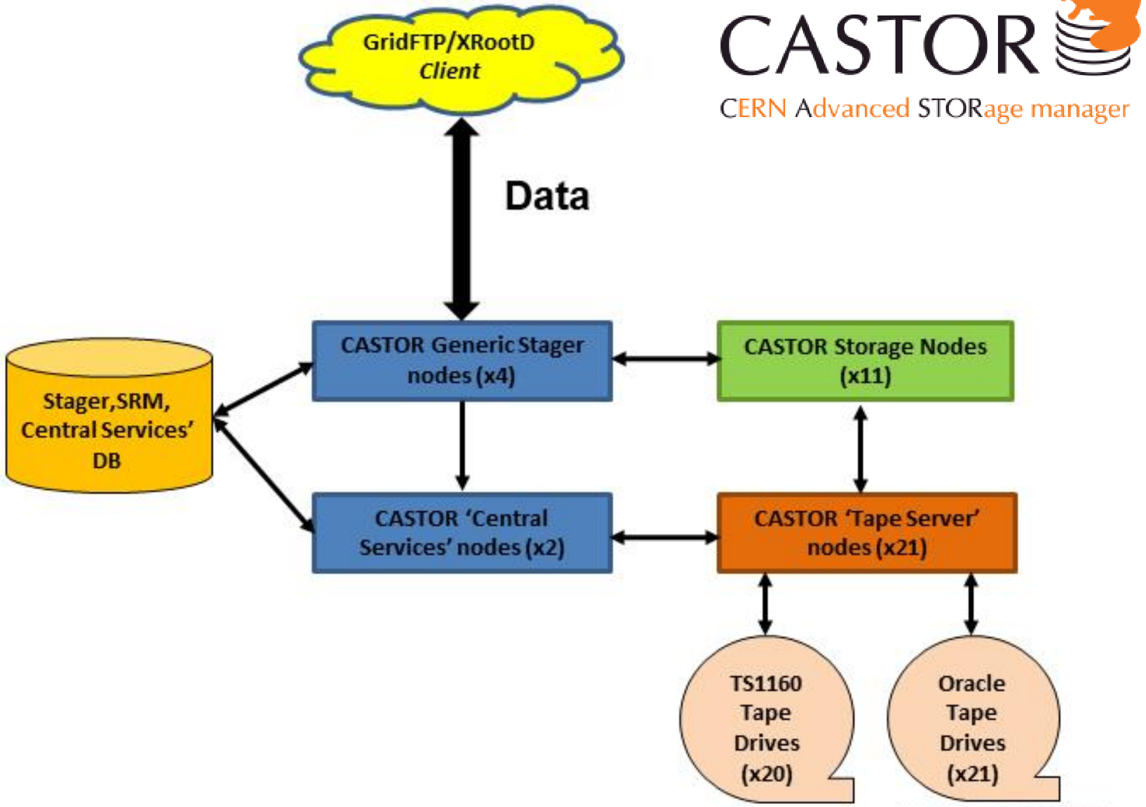
Outline

- Background (motivation, procurement exercise)
- Hardware (architecture, node set up)
- EOSCTA Progress
- Data transfer routes (CERN to RAL)
- Future plans

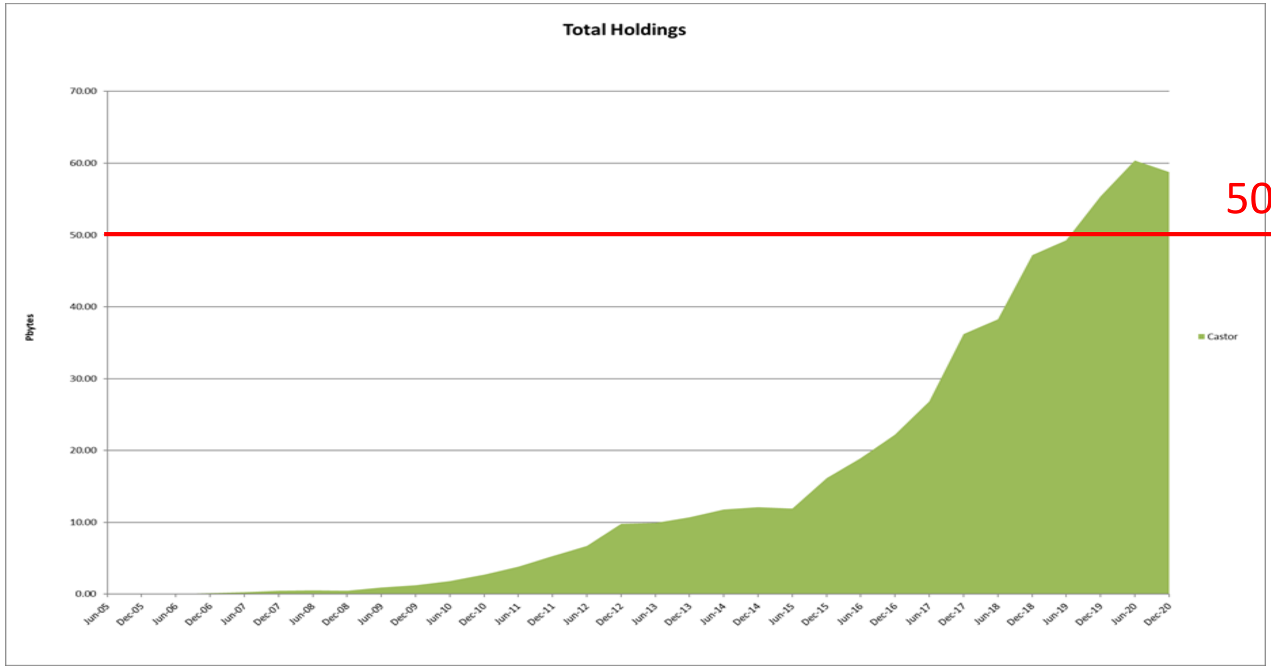
Motivation

- CASTOR has provided the tape archive service at RAL since 2006
- Designed and maintained by CERN who have now migrated to CTA
- RAL need to find a new solution or continue running CASTOR
- Opportunity to build a new service more aligned with STFC strategic objectives

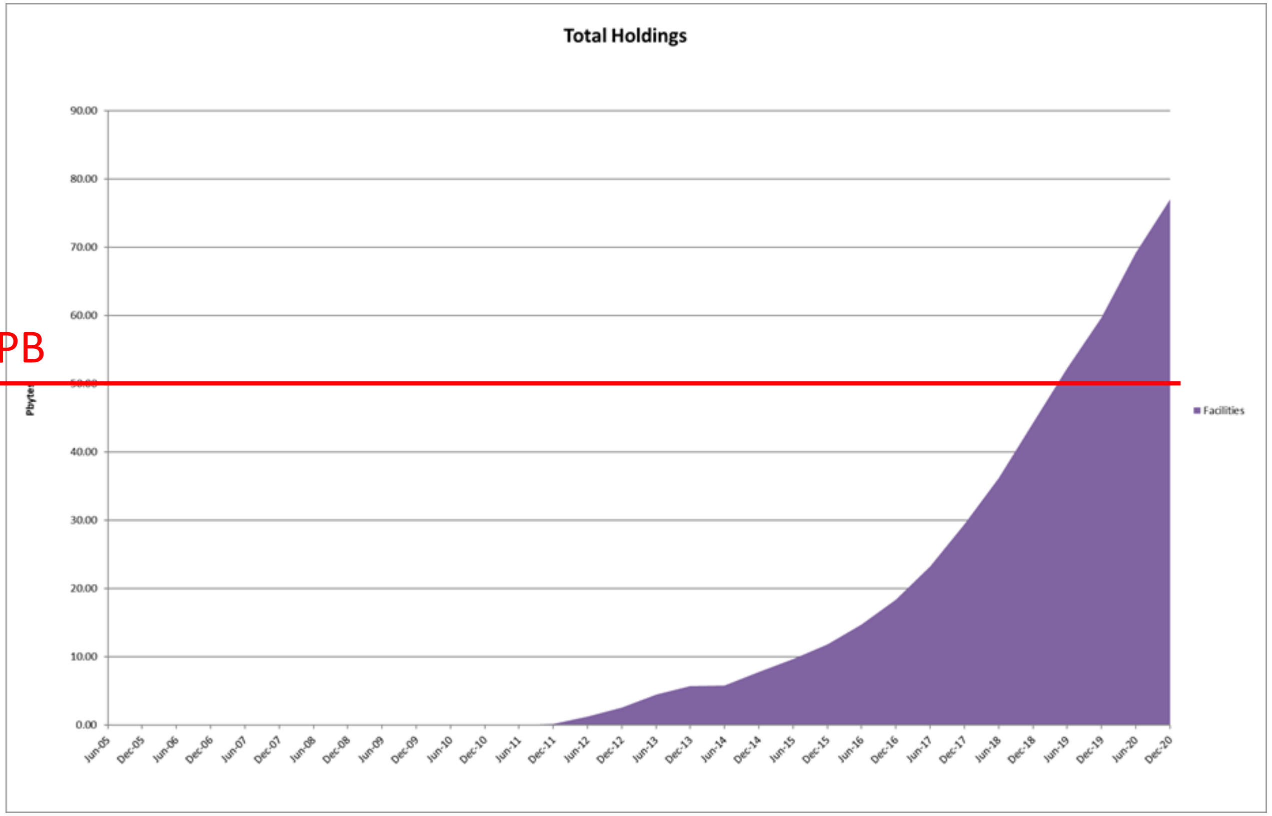
CASTOR
CERN Advanced STORage manager


GridFTP/XRootD
Client
Data
Stager,SRM, Central Services' DB
CASTOR Generic Stager nodes (x4)
CASTOR Storage Nodes (x11)
CASTOR 'Central Services' nodes (x2)
CASTOR 'Tape Server' nodes (x21)
TS1160 Tape Drives (x20)
Oracle Tape Drives (x21)

STFC Tier-1/Facilities Castor Data Volumes



50PB



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

UK RI Science and Technology Facilities Council

Tape Archive Solutions

With Castor EOL, RAL evaluated various options for replacement, including commercial products and CTA

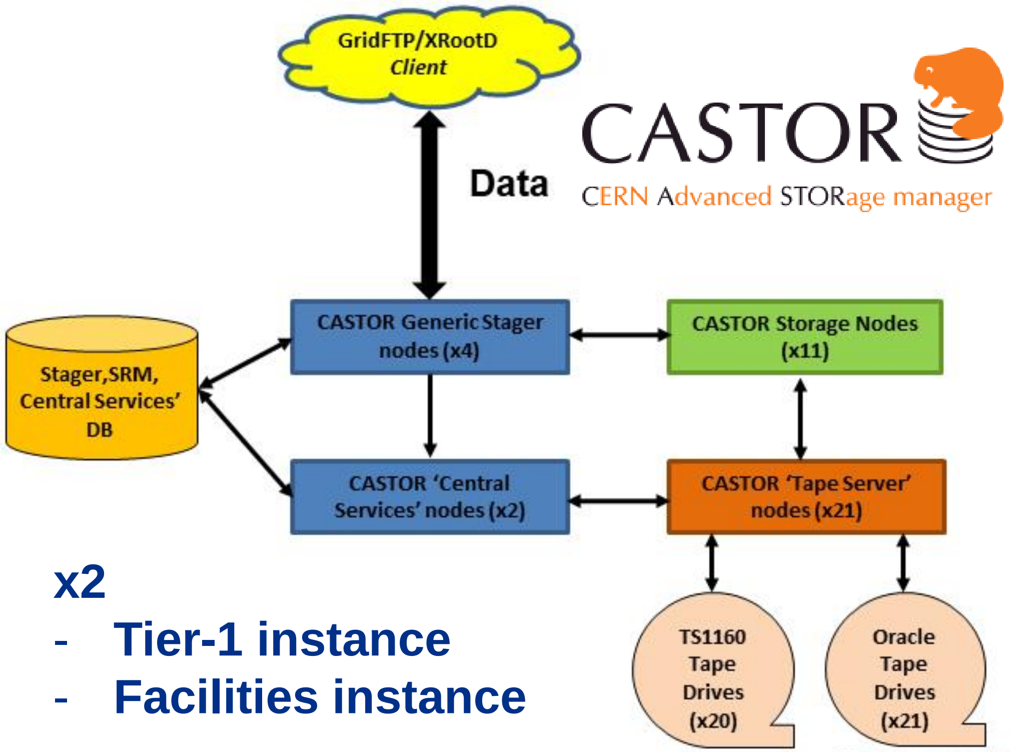
Commercial - several areas which made this less attractive/feasible option at RAL:

- Funding models
- Vendor lock-in
- Integration effort with our existing software
- Timescales to migrate from Castor

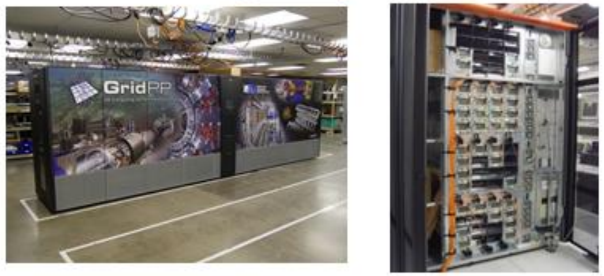
CTA provides significant advantages over other solutions for RAL:

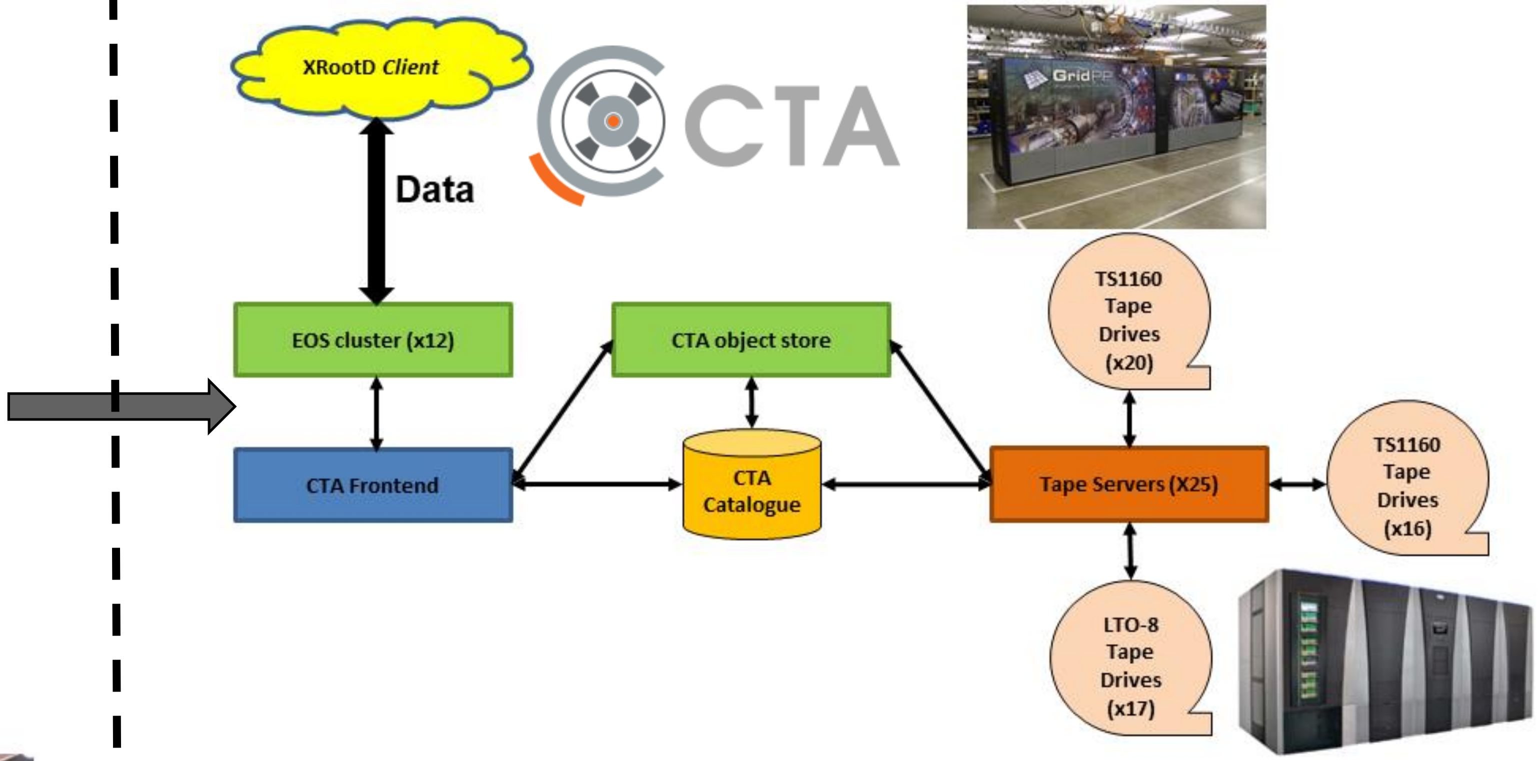
- Staff familiarity with many concepts from Castor and good collaboration with colleagues at CERN for many years
- Migration of data in situ
- Opportunity to move away from Oracle database software
- Opportunity to create a more unified service across STFC – one instance for Tier-1 and Facilities

From CASTOR to EOSCTA at RAL


GridFTP/XRootD
Client
Data
CASTOR
CERN Advanced STORage manager
Stager,SRM, Central Services' DB
CASTOR Generic Stager nodes (x4)
CASTOR Storage Nodes (x11)
CASTOR 'Central Services' nodes (x2)
CASTOR 'Tape Server' nodes (x21)
TS1160 Tape Drives (x20)
Oracle Tape Drives (x21)

x2
- Tier-1 instance
- Facilities instance





Transition to a simpler and more unified tape archive service able to cope with the LHC Run 3 data rates

UKRI Science and Technology Facilities Council

Hardware

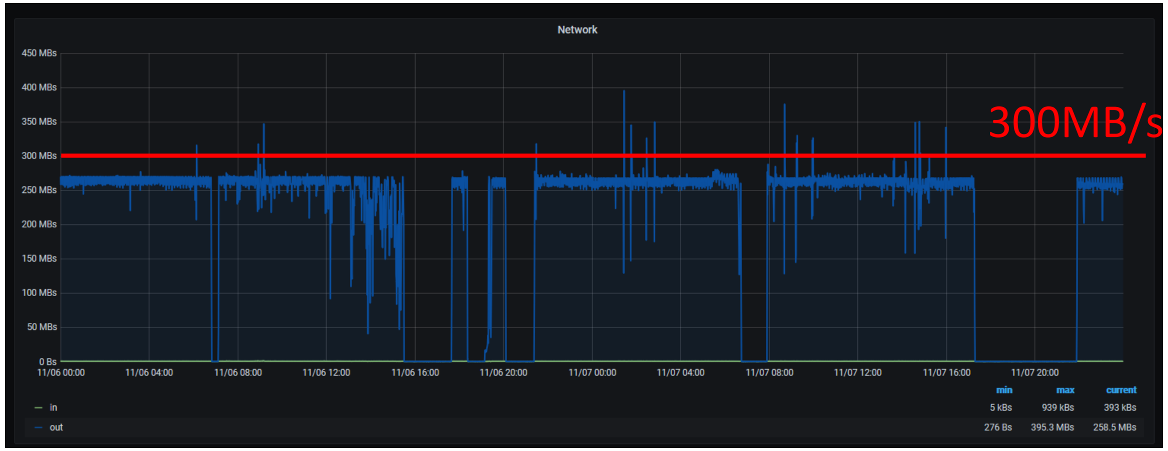
Node Type & Number	Function	Model	CPU	Memory	Disk	Network
EOS 12 x production 2 x test	Namespace management & disk cache	DELL R740XD	2 x Intel Xeon Gold 5218	192 GB	System + 1 NVMe + 16 x 2TB SSD	1 x Mellanox ConnectX-4 LX Dual Port 10/25GbE 1 x Intel Ethernet I350 Dual Port 1GbE BASE-T Adapter
Ceph 3 x production 2 x standby/dev	For transient data, queues and requests stored as objects in key-value store	DELL R6415	1 x AMD EPYC 7551	128GB	System + 8 x 4TB SSD	1 x Mellanox ConnectX-4 LX Dual Port 10/25GbE
Database 2 x Oracle RAC production 2 x Oracle RAC test	CTA catalogue	DELL PowerEdge R440	2 x Intel Xeon Gold 5222	192 GB	System + separate storage array (~90TB capacity)	1 x Broadcom 5720 Dual Port 1 GbE 1 x Dual-Port 1GbE On-Board LOM
Tape Server	RAL intend to allocate 1 tape server per 2 tape drives (initially)	DELL PowerEdge R640	2 x Intel Xeon Silver 4214	96 GB	2 x 240GB SSD SATA	1 x Mellanox ConnectX-4 LX Dual Port 10/25GbE
Frontend Servers (virtual)	Accepts archive/retrieve requests from EOS and send to CTA object store. Used for admin commands					

Tape Library Migration

- Support for Oracle tape ends mid-2020s
- Two Spectra TFinity libraries purchased in 2019 and 2020
- CTA is integrated with Spectra and IBM currently, but not Oracle
- Migrate 130PB of data from Oracle SL8500 to Spectra before CTA goes into prod
- Set up a separate Tier-1 Repack CASTOR instance
  - ➢ Single generic CASTOR headnode (stager): 6 CPU cores and 32GB RAM
  - ➢ Tape buffer: 9 x HDD and 4 x SSD disk servers → 625TB
  - ➢ Initial drive allocation: 10 x T10KD for reading (250MB/s) and 6 x TS1160 for writing (400MB/s)
  - ➢ Final drive allocation: 14 x T10KD for reading and 8 x TS1160 for writing (accelerated rate)

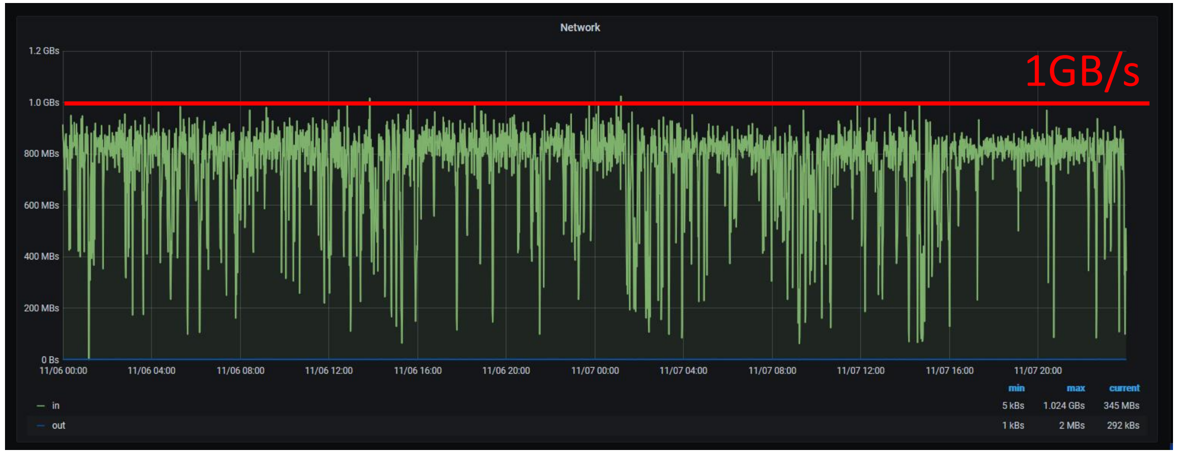
Tape Library Migration

Reading rate from one T10KD drive (Oracle)

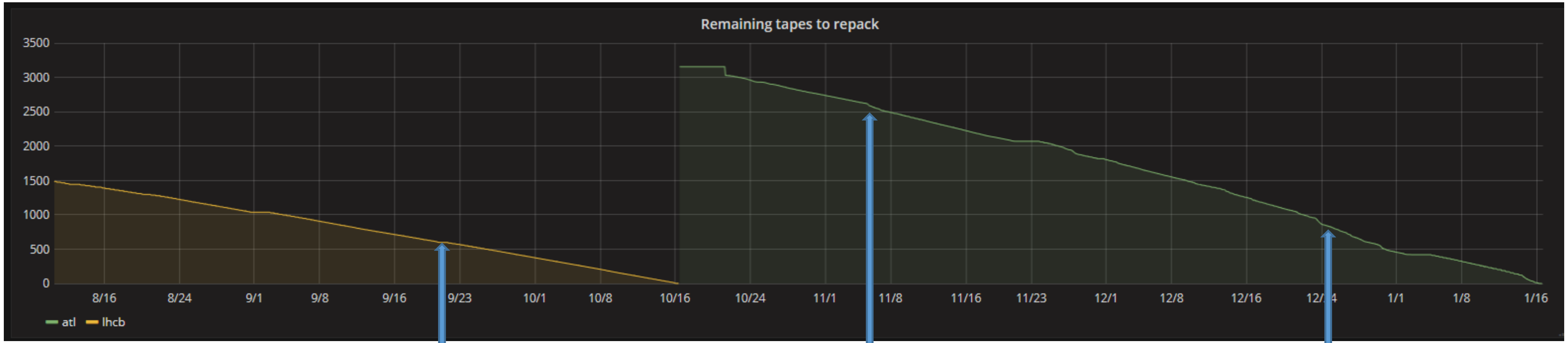


	min	max	current
in	5 kBs	939 kBs	393 kBs
out	276 Bs	395.3 MBs	258.5 MBs

Writing rate to two TS1160 drives (Spectra)



	min	max	current
in	5 kBs	1.024 GBs	345 MBs
out	1 kBs	2 MBs	292 kBs



~20 tapes/day

Allocate more drives
for repacking

~25 tapes/day

Science and
Technology
Facilities Council

Tier-1 Database Clean-up

- CASTOR was used to manage also disk-only data
- ATLAS Rucio naming convention creates a lot of directories
- Remove old disk-only directories no longer required
- Iterative procedure of directory & file search and deletion
- Result: smaller DB schemas to migrate to CTA
- Deleted entries per VO:

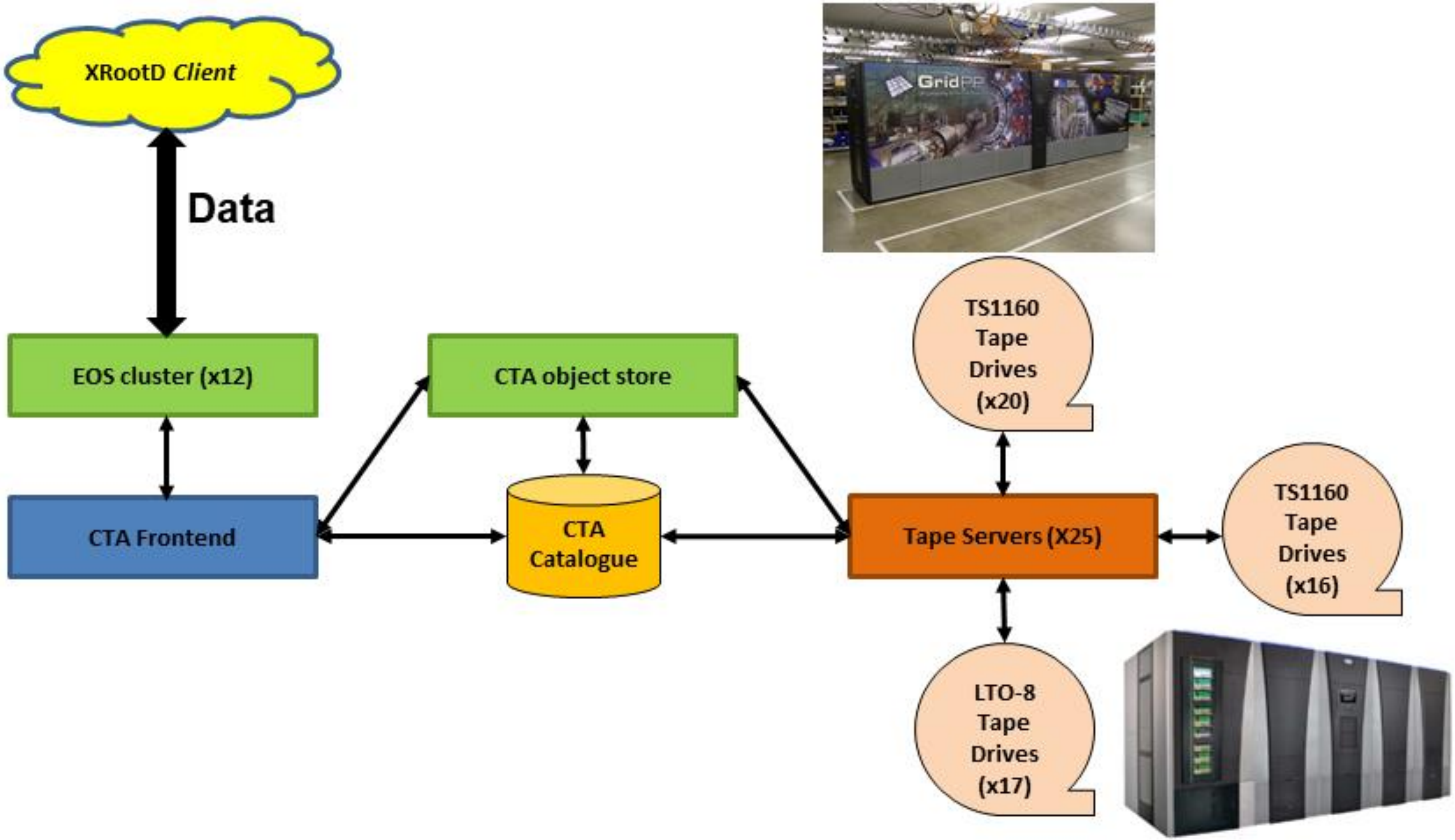
ATLAS:	20,133,182
CMS:	973,006
LHCb:	8,543,377

Total: 29,649,565

36% of the namespace size before deletion!

Science and Technology Facilities Council

CTA design at RAL



CTA Tape deployment

- CTA tape system resembles CASTOR from which it has evolved
  - ✓ Storage Class (CTA) ↔ File Class (CASTOR)
  - ✓ Archive Route (CTA) ↔ Migration Route (CASTOR)
  - ✓ Requester mount rule (CTA) ↔ Recall group (CASTOR)
  - ✓ Tape pools (CTA) ↔ Tape pools (CASTOR)
- A CTA tape server is (basically) a CASTOR tape server

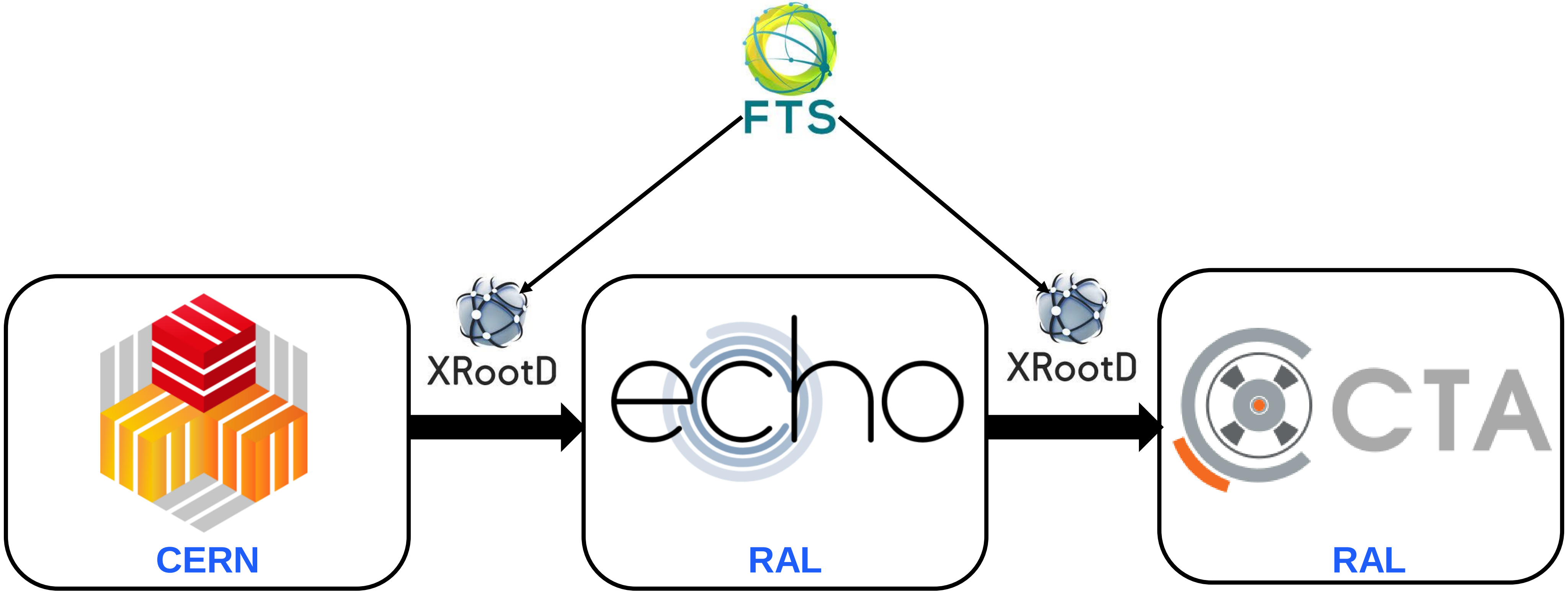
CTA Database deployment

- During development we are using PostgreSQL for the CTA catalogue and mhVTL as the virtual tape library.
- When we migrate from Castor to CTA, we will use Oracle and follow the CERN migration path.
  - In the longer term we aim to migrate from Oracle to PostgreSQL
- CTA object store will be provisioned by Ceph, which we have substantial experience with

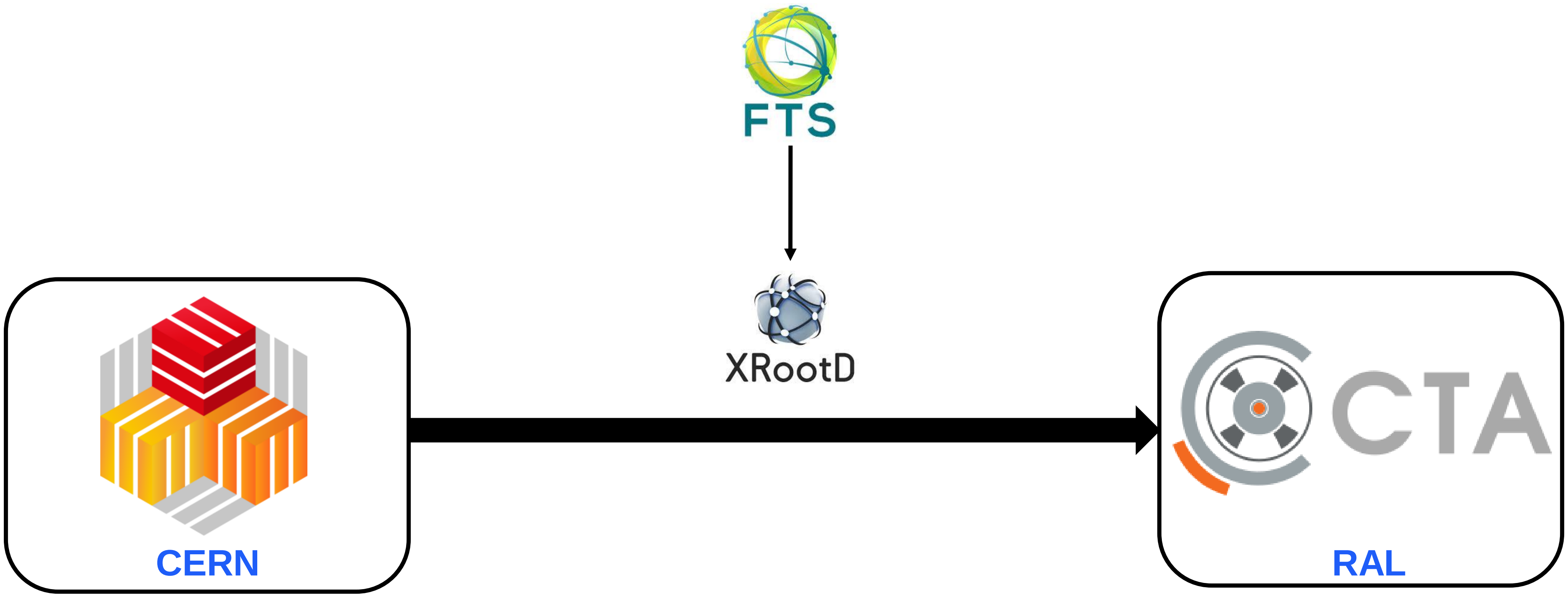
EOS-CTA deployment

- No prior experience in running EOS at RAL
  - Especially an all SSD EOS instance requiring different performance tunings
  - Work in progress to understand the set up of an EOS instance
- CERN advised us to set up a K8s EOS-CTA instance on a standalone VM
  - This was very educational
  - We also tried creating a Docker EOS instance
- Currently we have a cluster of cloud VMs running EOS MGM/FST/QuarkDB
- Delayed access to hardware due to Covid-19 disruption

Data transfer routes 1



Data transfer routes 2



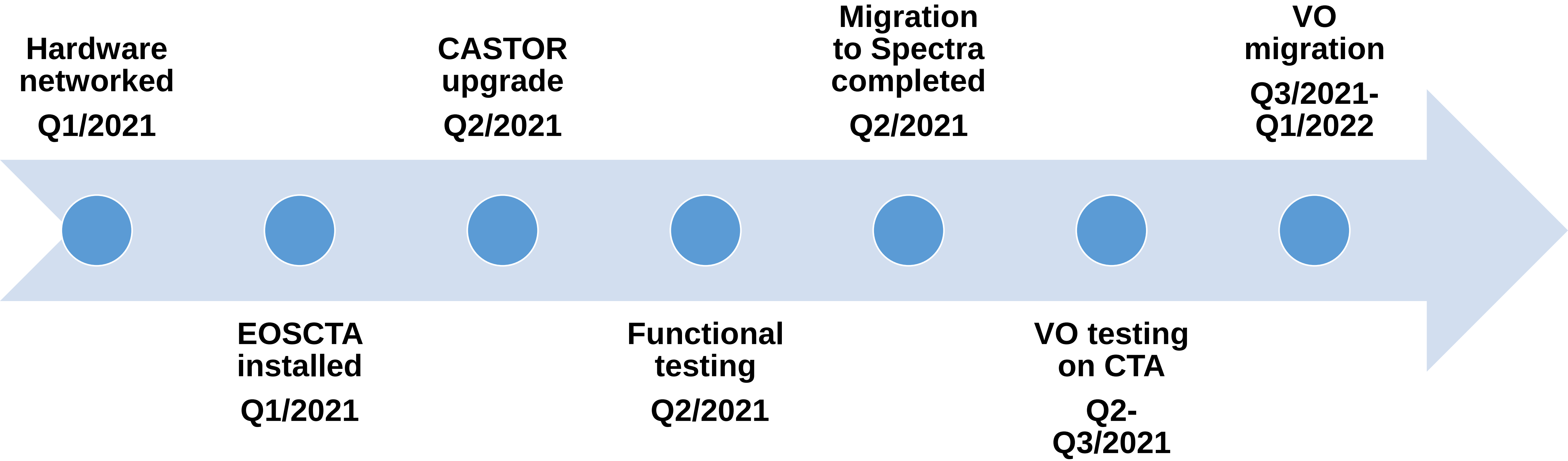
For RAW data export, probably best to transfer directly to RAL CTA.

For other types of transfer, an FTS multi-hop via Echo may be better.

We will need to test out both.

Science and
Technology
Facilities Council

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

THANK YOU!

Science and
Technology
Facilities Council