UKRI
Science and
Technology
Facilities Council

Tier-1 Strategy

Alastair Dewhurst

Agenda

1 Introduction

2 Business as Usual + Continuous Improvement

3 Antares (CTA)

4 Network + Core infrastructure Improvements

5 Rucio and Data Lake development

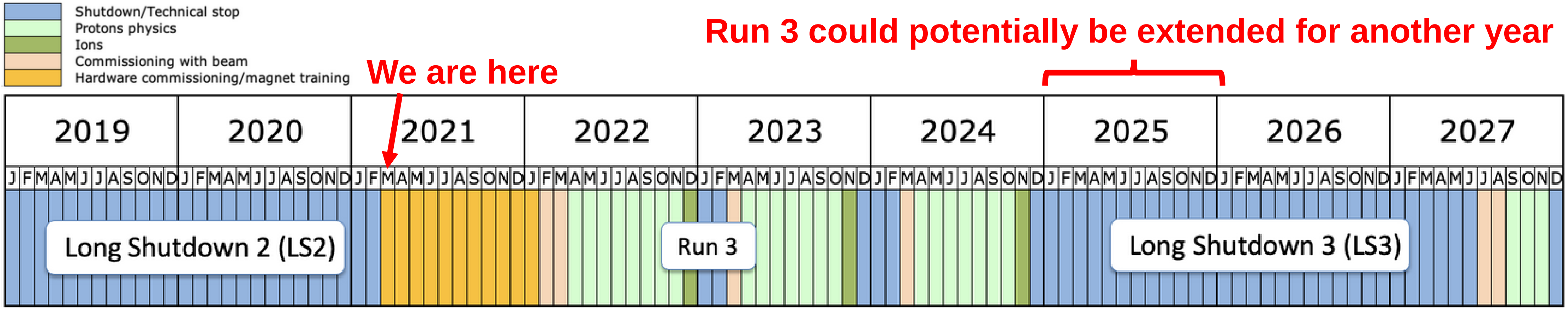
6 Authentication and Authorization

Image © STFC Alan Ford

Science and Technology Facilities Council

Introduction

- On 9th March 2020 I gave a presentation regarding the Tier-1’s plan for GridPP6.
  - I followed this up with a presentation of 12th October as the pandemic had meant a lot of things had changed.
- This talk focuses on what I would like to achieve over the coming year to ensure we are ready for LHC Run 3.



We are here

Run 3 could potentially be extended for another year

GridPP Strategic Goals

1. *To deliver STFC's MoU commitment to CERN and the WLCG by ensuring that GridPP meets the challenge of higher data rates and data volumes of LHC Run 3.*
2. *To prepare for the 2027 start of HL-LHC (LHC Run 4) by influencing WLCG's future technical direction and contributing to development.*
3. *To provide broader benefit to STFC and their communities by continuing established initiatives to reduce the operational cost of the infrastructure, whilst increasing support for non-LHC communities and developing common infrastructure and operations.*

GridPP Strategic Goals

1. *To deliver STFC's MoU commitment to CERN and the WLCG by ensuring that GridPP meets the challenge of higher data rates and data volumes of LHC Run 3.*
2. ***To prepare for the 2027 start of HL-LHC (LHC Run 4) by influencing WLCG's future technical direction and contributing to development.***
3. *To provide broader benefit to STFC and their communities by continuing established initiatives to reduce the operational cost of the infrastructure, whilst increasing support for non-LHC communities and developing common infrastructure and operations.*

GridPP Strategic Goals

1. *To deliver STFC's MoU commitment to CERN and the WLCG by ensuring that GridPP meets the challenge of higher data rates and data volumes of LHC Run 3.*
2. *To prepare for the 2027 start of HL-LHC (LHC Run 4) by influencing WLCG's future technical direction and contributing to development.*
3. ***To provide broader benefit to STFC and their communities by continuing established initiatives to reduce the operational cost of the infrastructure, whilst increasing support for non-LHC communities and developing common infrastructure and operations.***

Science and
Technology
Facilities Council

Business as Usual + Continuous Improvement

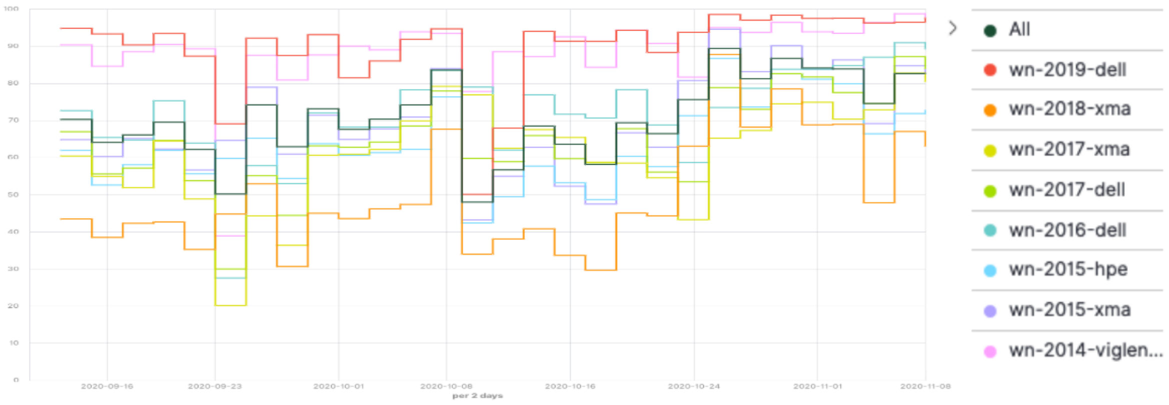
LHC Resource Requirements

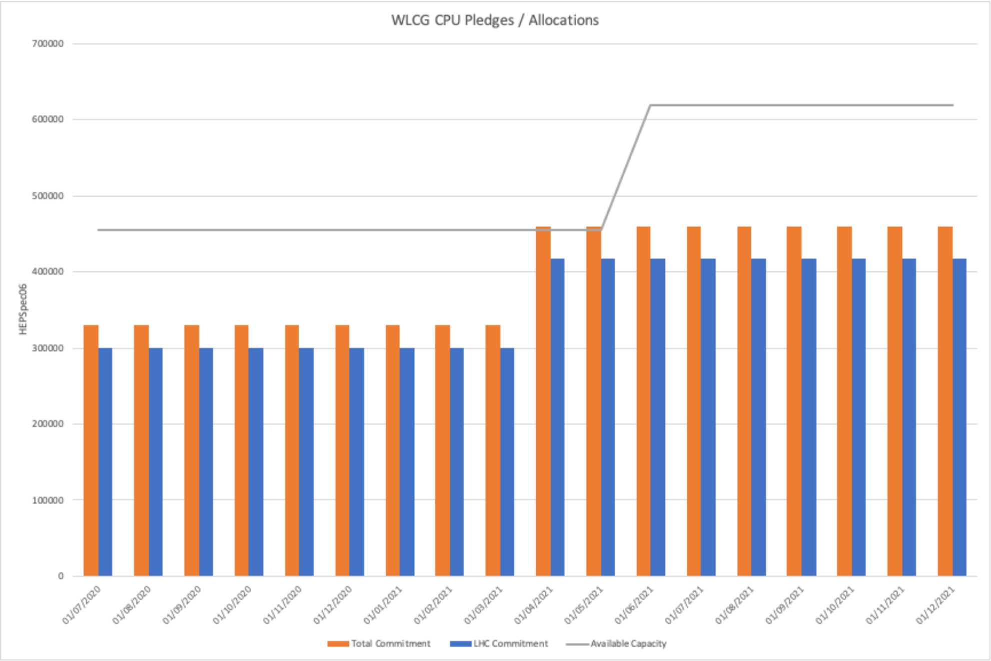
VO	FY20/21 Tape (TB)	FY21/22 Tape (TB)	FY20/21 Disk (TB)	FY21/22 Disk (TB)	FY20/21 CPU (HS06)	FY21/22 CPU (HS06)
ALICE	1,131	1,710	1,320	1,599	10,950	14,940
ATLAS	32,708	34,780	13,024	15,540	159,692	173,160
CMS	17,600	17,600	5,440	5,440	52,000	52,000
LHCb	15,270	17,078	8,370	8,460	73,800	129,150
LHC Total	**66,709**	**71,168**	**28,154**	**31,039**	**303,942**	**369,250**

- Formal LHC resource request were agreed in September.
  - Due to Covid-19 late deployment (after April 2021) of resources is understandable.
- Modest increases in most requests
  - Large LHCb CPU usage increase, due to a one-off correction in their CPU calculation for one workflow.

CPU deployment

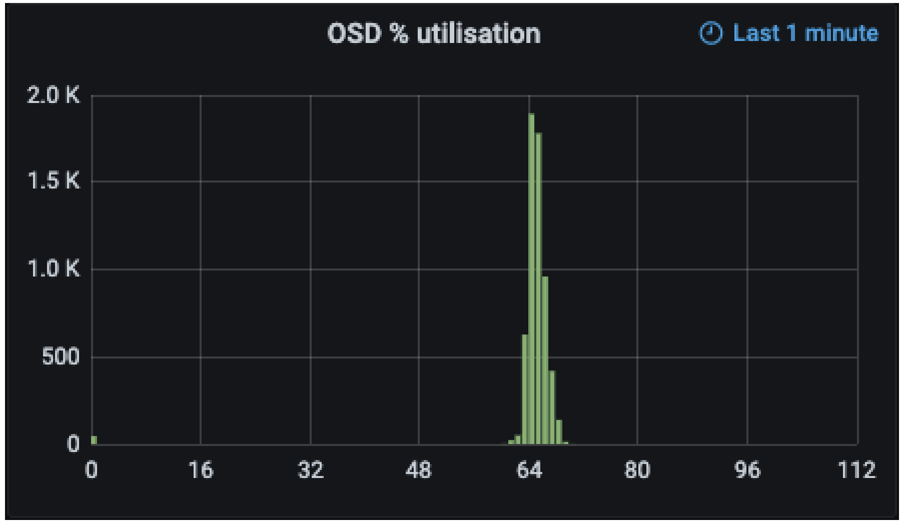
- For CPU we can just meet our allocation without deploying the XMA20 CPU.
- We want to put them in soon as SSD backed machines have better efficiency.

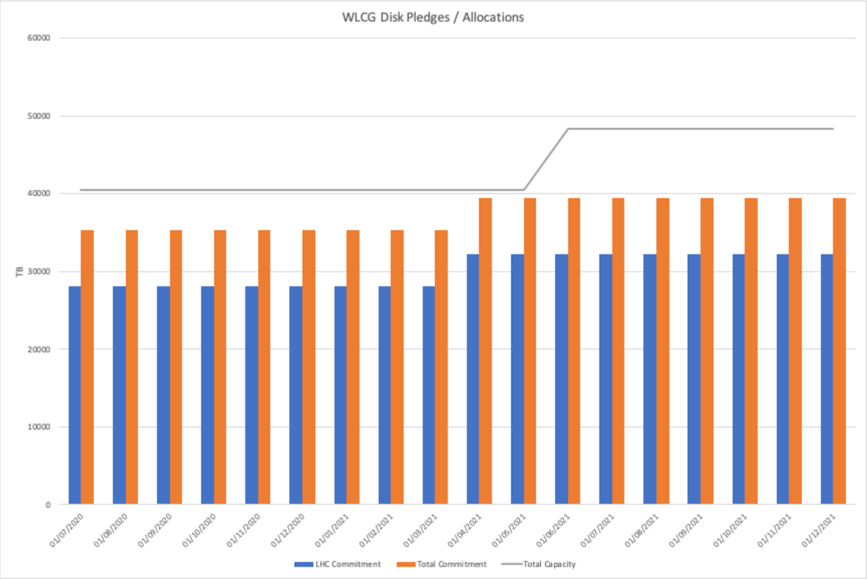




Disk deployment

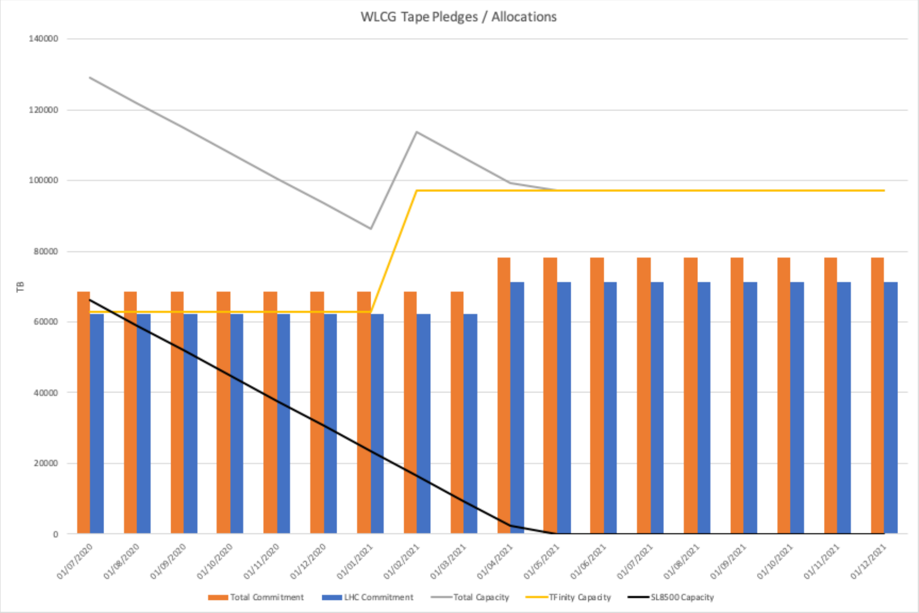
- For Disk we can meet our allocation without deploying the XMA20 CPU.
  - The Dell 19 isn't properly deployed yet,
- Echo has plenty of spare capacity currently as many non-LHC VOs are not using their allocation.



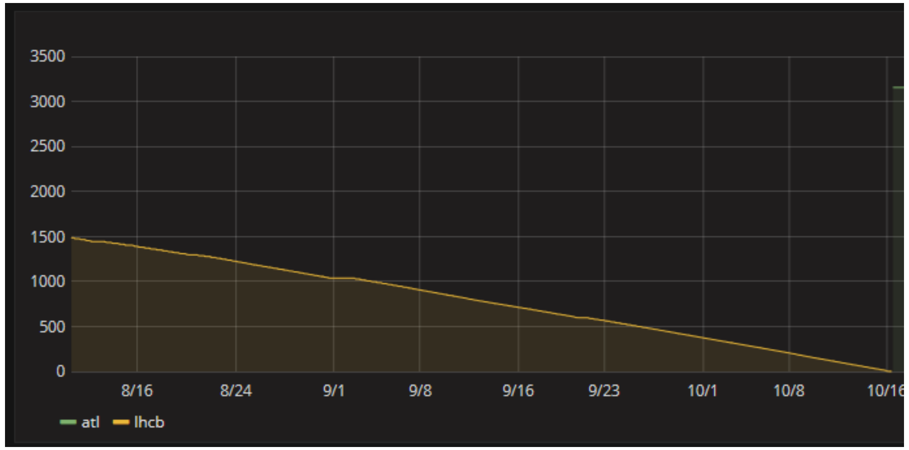


Tape Deployment

- Migration to Tfinity Tape Robot is going very well!
  - 1317 (CMS) tapes left to do ~ 7 weeks.
- We purchased 79.2PB of tape media this year, which should last us until Mid 2022.
  - Next purchase will hopefully be new technology.



~1.3PB migrated a week.

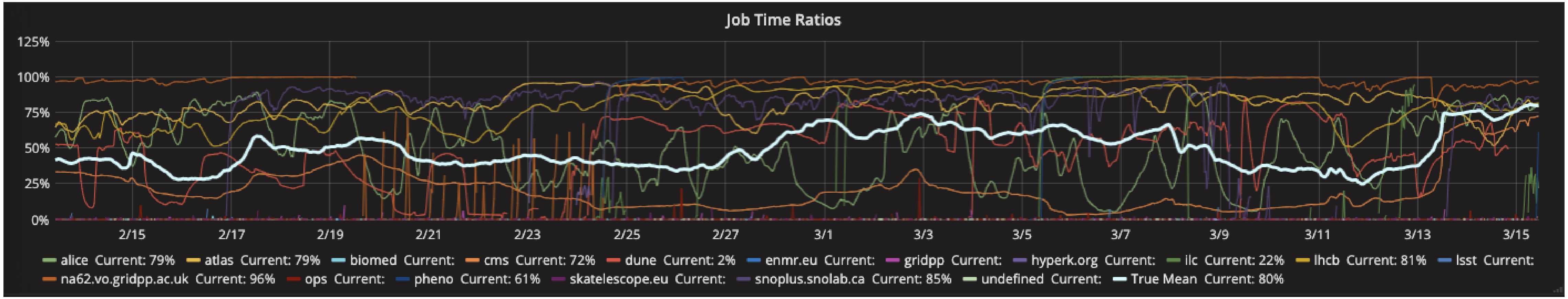


Lots of XRootD

- For at least the duration of Run 3 the LHC experiments will use XRootD to access their data for jobs.
  - Will arrange an XRootD tutorial/workshop for Q2.
- With the retirement of GridFTP, Echo will become an XRootD endpoint for the LHC Experiments.
  - XRootD will perform http third party copies (via a plugin).
  - XRootD will also support the new authentication mechanism (tokens)

Batch Farm

- Two major problem affecting Job Efficiency:
  - Lack of Vector Reads
  - WN access external data via firewall.
- Both of these problems should be resolved in the next 3 months.
  - There will be smaller problems we need to fix.



Monitoring and Operations

- A significant amount of effort has gone in to modernized and consolidate monitoring services over the last year.
- Nagios has been migrated to ICINGA
- influxdb has been expanded and moved to SSD storage.
- We are starting to move our oncall system to Opsgenie
- We would like to move to using JIRA for our ticket system
- elasticsearch is also an important tool that we can hopefully improve.

FY20/21 Capital and Resource Spend

Item	Cost (inc VAT)	Description
CPU	£695,677.25	112 WN (14336 cores)
Disk	£607,935.74	48 Storage Nodes (18432 TB)
Tape Media	£949,459.82	79.2PB
Network	£891,035.58	LHCOPN Upgrade + Maintenance, New Core Network, contractor
Other	~£200,000.00	
Total Spend	**~£3,350,000.00**	

FY21/22 Procurement

- It is expected that funds will be a lot tighter next year.
  - £1.3 million Capital expected although we still need to plan for opportunistic Capital.
  - Resource spending is reducing, and we almost certainly won't get as much of the travel budget.
- We do not need to buy any more tape media.
- CPU and Disk purchases similar to this year.
  - Trying to keep increment to Echo similar each year.
- Purchase of Tau server to test new Echo setup.
- Paying XMA / Boston to benchmark CPU

Science and
Technology
Facilities Council

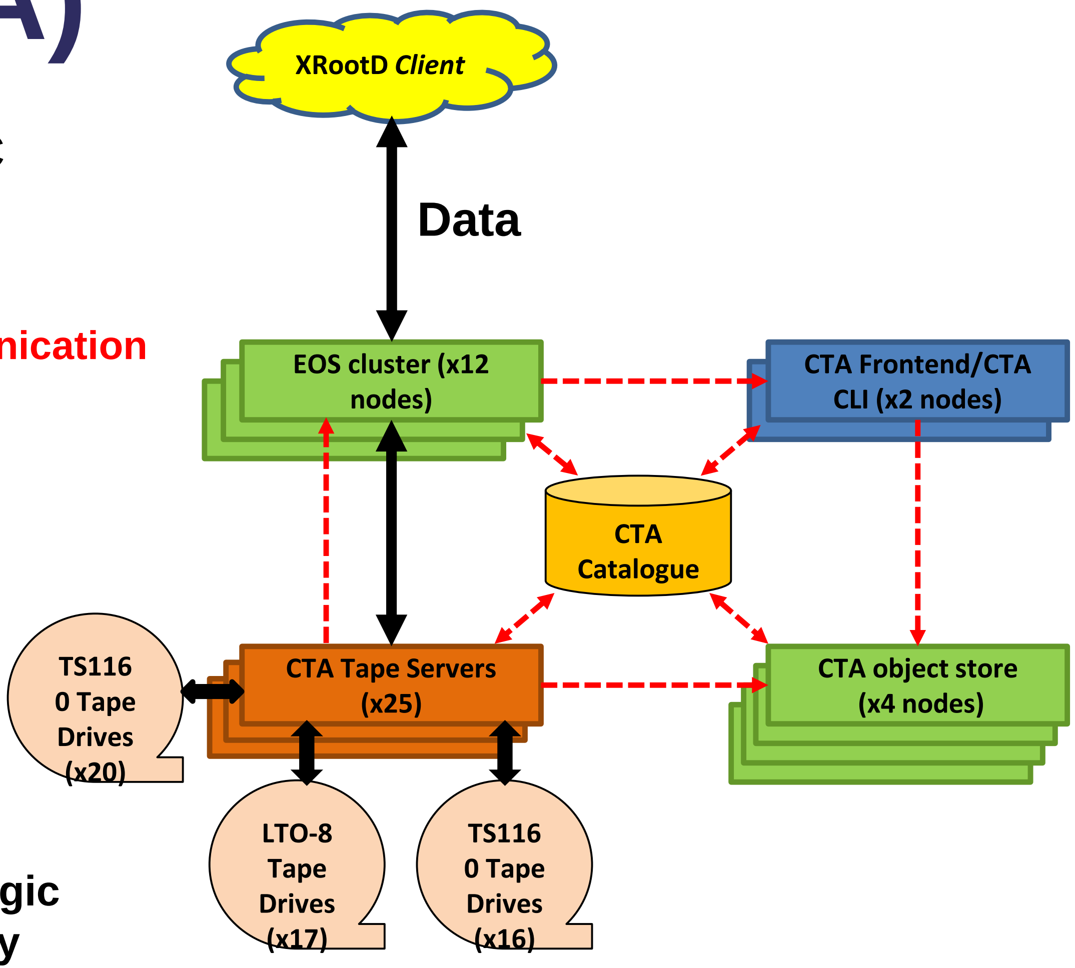
Antares (CTA)

Antares (CTA)

A New Tape ARchivE for STFC

Communication

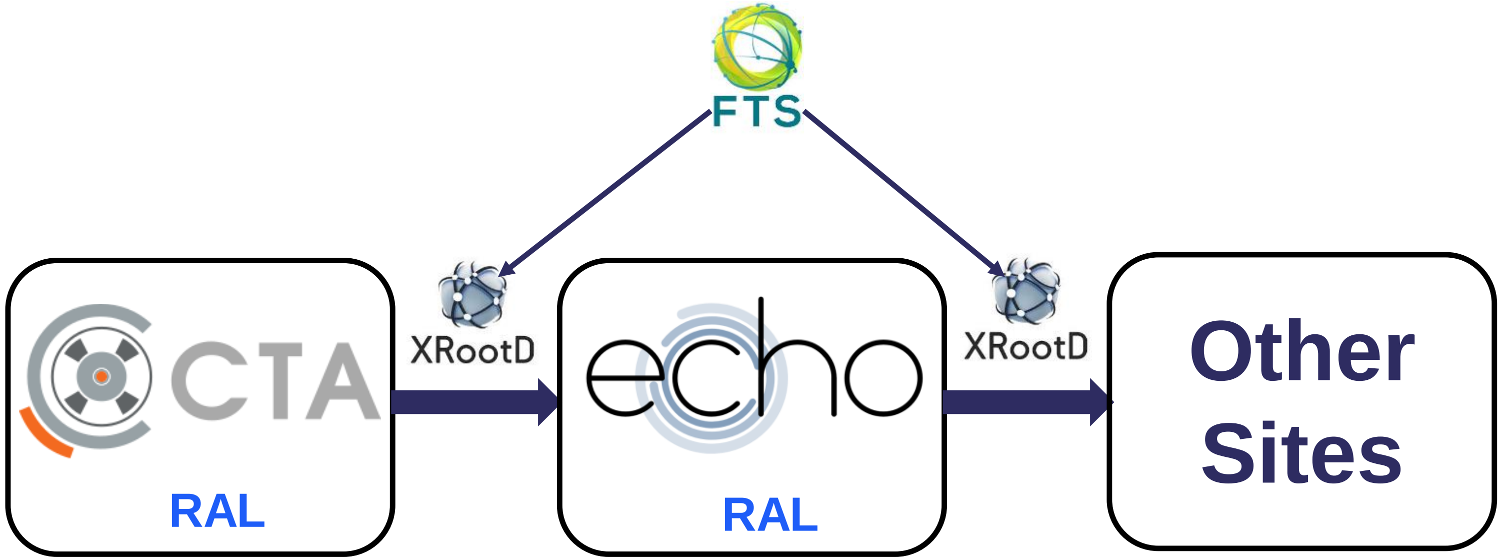
Data


XRootD Client
Data
EOS cluster (x12 nodes)
CTA Frontend/CTA CLI (x2 nodes)
CTA Catalogue
TS116 0 Tape Drives (x20)
CTA Tape Servers (x25)
CTA object store (x4 nodes)
LTO-8 Tape Drives (x17)
TS116 0 Tape Drives (x16)

Tier-1 Spectra Logic Tfinity tape library

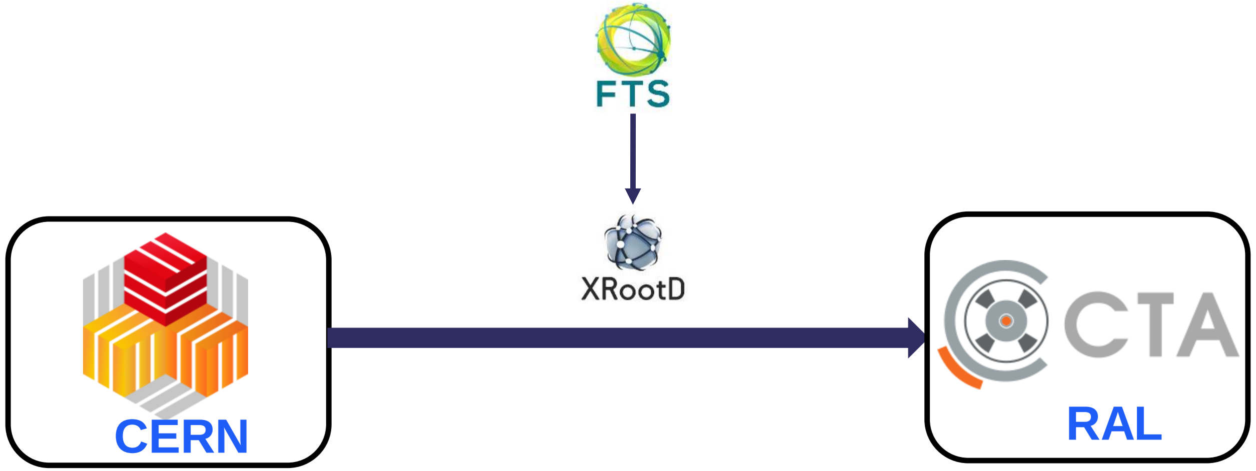
Facilities SpectraTfinity tape library

Data Transfer Routes?



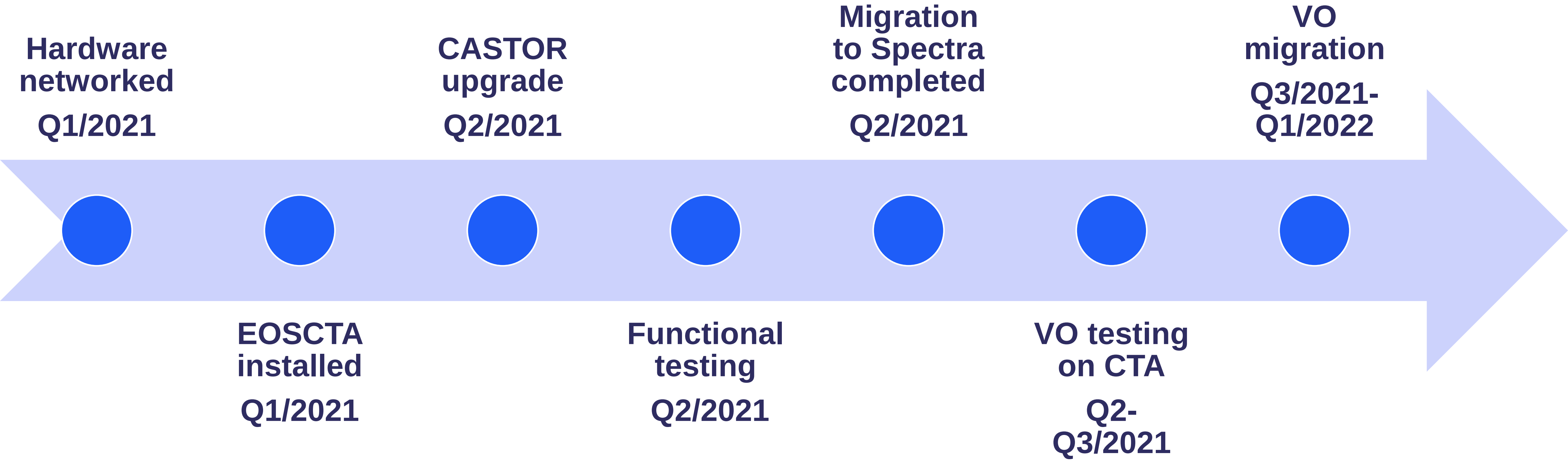
For other types of transfer, an FTS multi-hop via Echo may be better.

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



Important for VO Liaisons to understand and be able to test use cases

Migration Plan



Hardware networked
Q1/2021

EOSCTA installed
Q1/2021

CASTOR upgrade
Q2/2021

Functional testing
Q2/2021

Migration to Spectra completed
Q2/2021

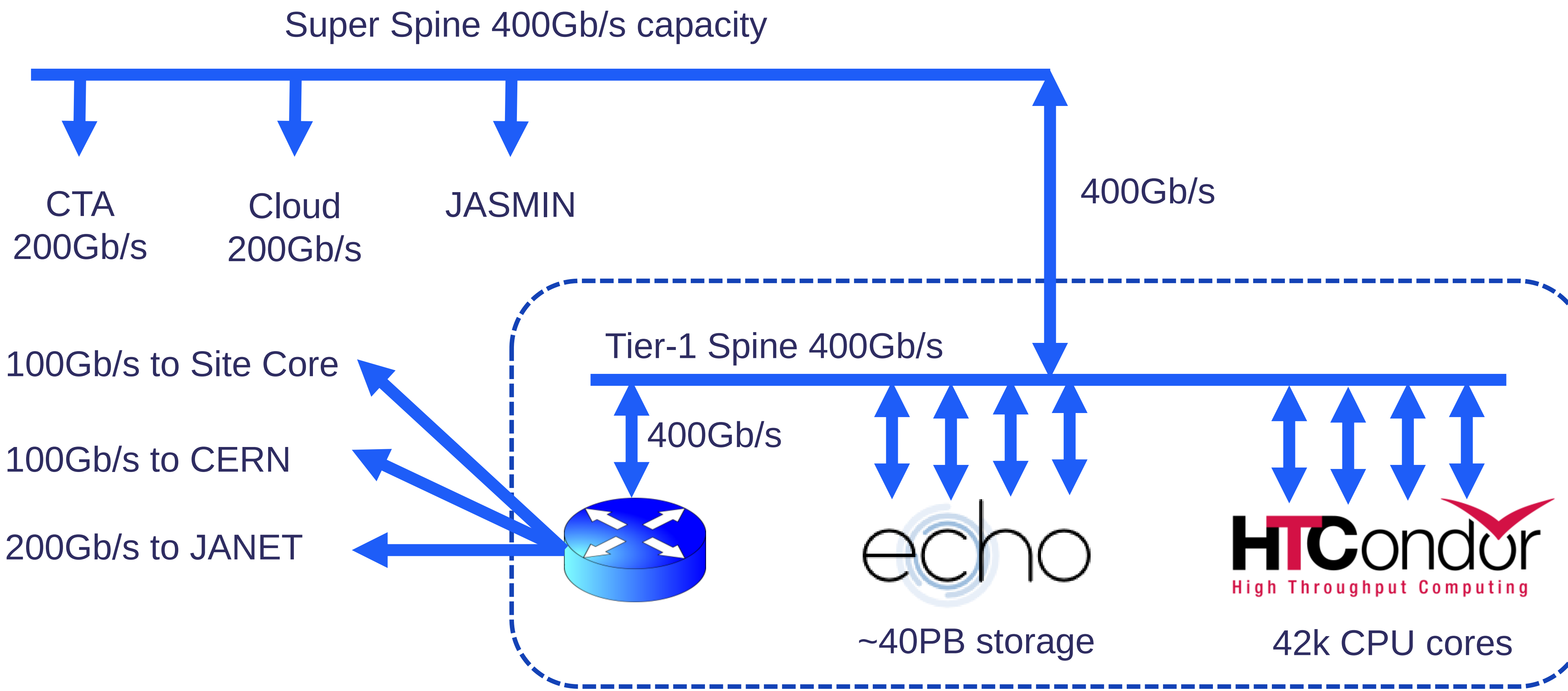
VO testing on CTA
Q2-Q3/2021

VO migration
Q3/2021-Q1/2022

Science and
Technology
Facilities Council

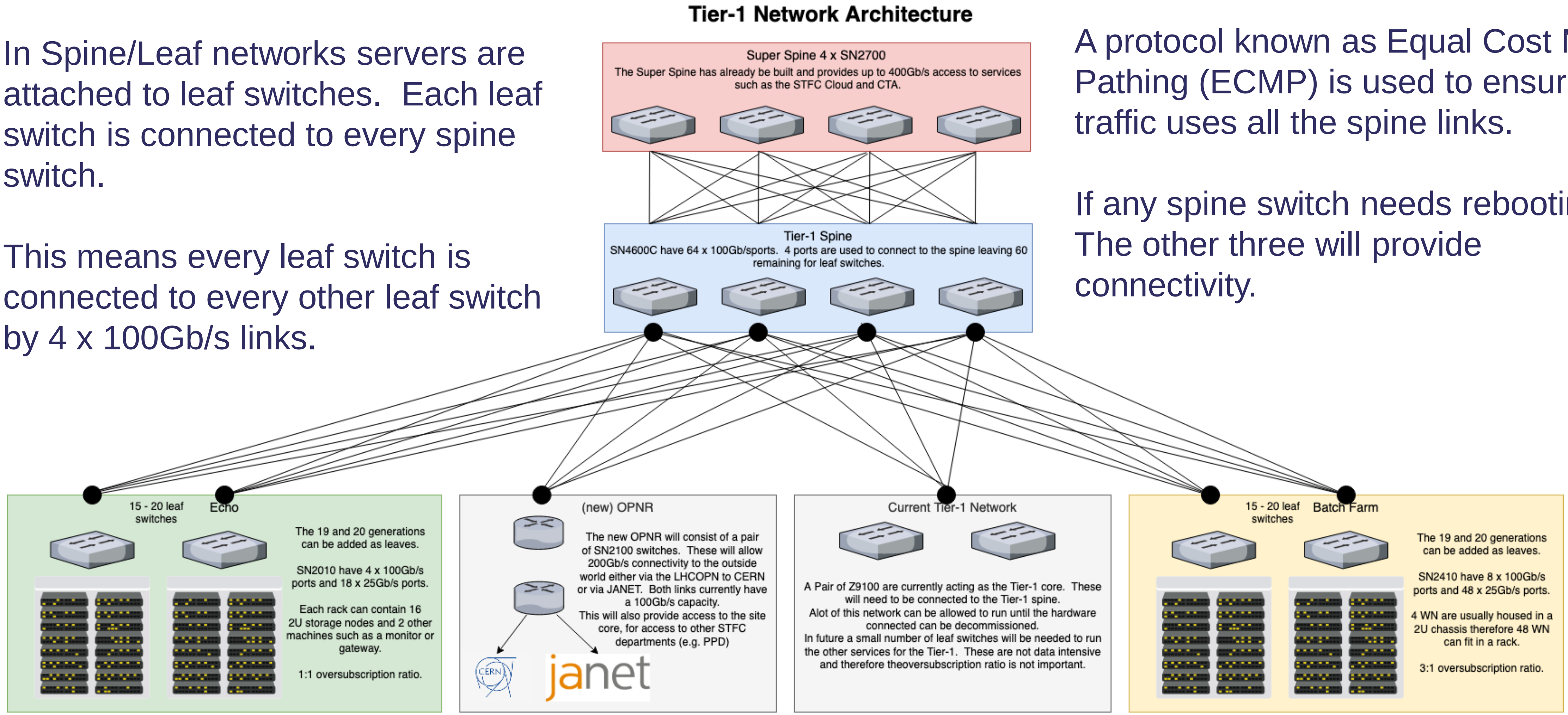
Network Evolution + Core infrastructure Improvements

Core network infrastructure



In Spine/Leaf networks servers are attached to leaf switches. Each leaf switch is connected to every spine switch.

This means every leaf switch is connected to every other leaf switch by 4 x 100Gb/s links.



Tier-1 Network Architecture

Super Spine 4 x SN2700

The Super Spine has already be built and provides up to 400Gb/s access to services such as the STFC Cloud and CTA.

Tier-1 Spine

SN4600C have 64 x 100Gb/sports. 4 ports are used to connect to the spine leaving 60 remaining for leaf switches.

Echo

15 - 20 leaf switches

The 19 and 20 generations can be added as leaves.

SN2010 have 4 x 100Gb/s ports and 18 x 25Gb/s ports.

Each rack can contain 16 2U storage nodes and 2 other machines such as a monitor or gateway.

1:1 oversubscription ratio.

(new) OPNR

The new OPNR will consist of a pair of SN2100 switches. These will allow 200Gb/s connectivity to the outside world either via the LHCOPN to CERN or via JANET. Both links currently have a 100Gb/s capacity.
This will also provide access to the site core, for access to other STFC departments (e.g. PPD)

Current Tier-1 Network

A Pair of Z9100 are currently acting as the Tier-1 core. These will need to be connected to the Tier-1 spine.
Alot of this network can be allowed to run until the hardware connected can be decommissioned.
In future a small number of leaf switches will be needed to run the other services for the Tier-1. These are not data intensive and therefore theoversubscription ratio is not important.

Batch Farm

15 - 20 leaf switches

The 19 and 20 generations can be added as leaves.

SN2410 have 8 x 100Gb/s ports and 48 x 25Gb/s ports.

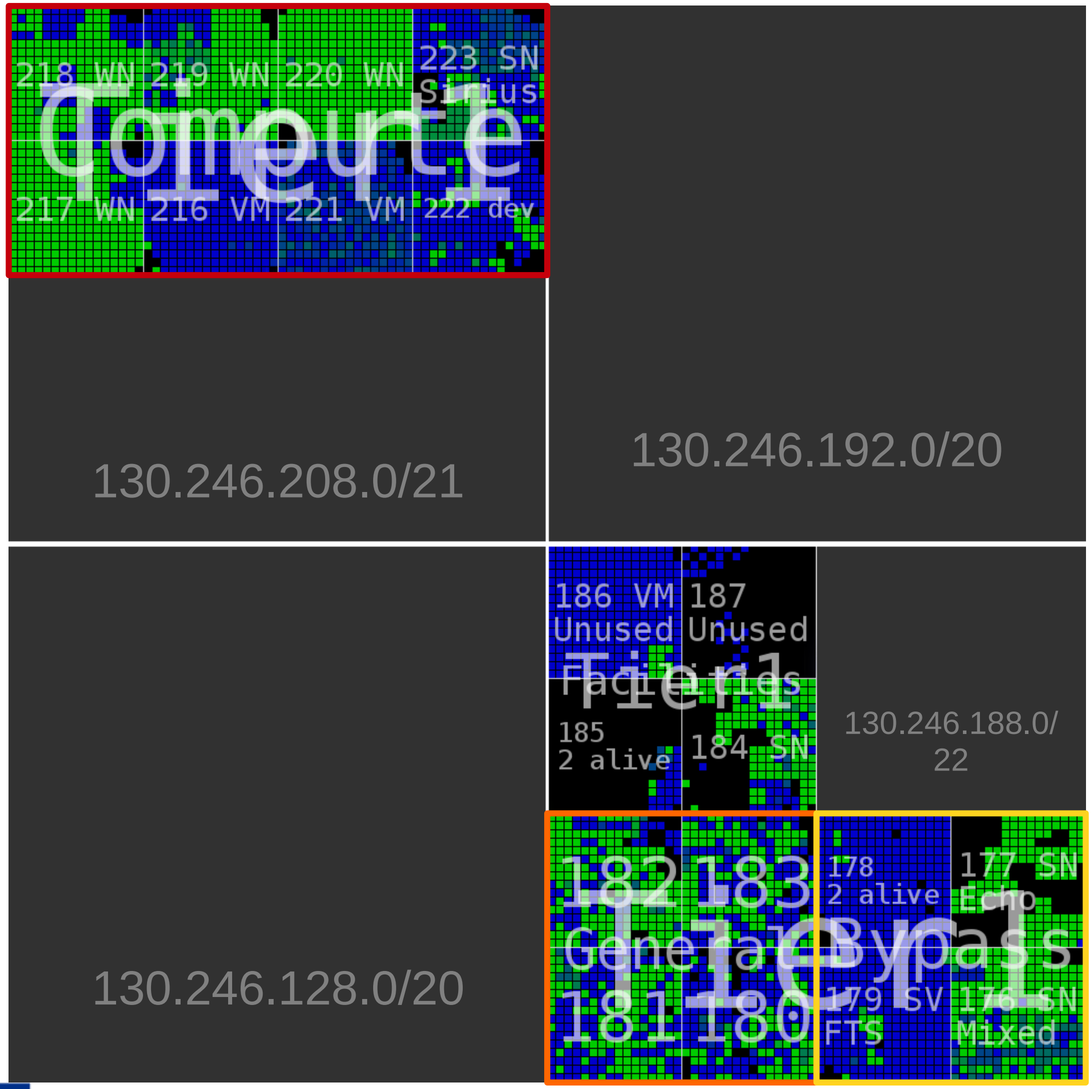
4 WN are usually housed in a 2U chassis therefore 48 WN can fit in a rack.

3:1 oversubscription ratio.

A protocol known as Equal Cost Multi Pathing (ECMP) is used to ensure traffic uses all the spine links.

If any spine switch needs rebooting, The other three will provide connectivity.

Tier-1 Subnets



3 x Tier 1 subnets:

OPN: 130.246.176.0/22

Services: 130.246.180.0/22

Compute: 130.246.216.0/21

Subnet design was done a decade before services like CMS AAA were thought of.

Time Line

1) Build new Tier-1 Network
   - XMA are doing all installation and cabling inside the data centre (today). Target completion April 1st.
   - Dedicated contractor effort (Anil) to configure setup. Target completion May 1st.
2) Connect Network pods to Super Spine
   - Can happen in parallel to 1). Target completion May 1st.
3) Switch Peering from OPNR to TIE Router.
4) Announce 130.246.216.0/21 and 2001:630:58:1820/64 to LHCOPN.
5) Announce 130.246.216.0/21 and 2001:630:58:1820/64 to LHCONE.
6) Migrate older hardware to new network. Q3 2021

Converged Infrastructure

- Converged infrastructure is:
  - *“A way of structuring an information technology system which groups multiple components into a single optimized computing package.”*
- I would like us to not put old hardware on the new infrastructure.
  - I will be buying ~10 standard machines which can be used for multiple things.
  - Echo Gateways & Monitors, InfluxDB and PostgreSQL servers all same underlying machine.

Science and
Technology
Facilities Council

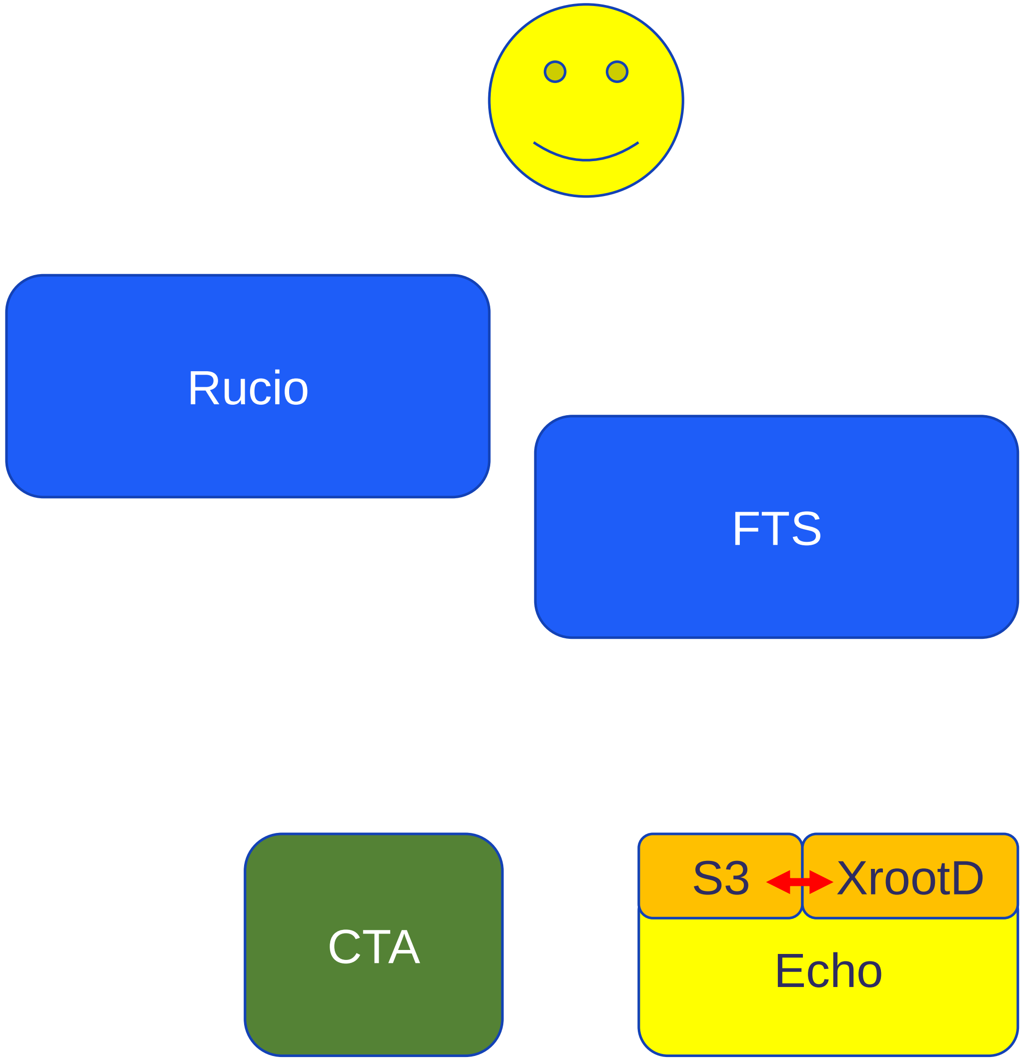
Rucio and Data Lake Development

ExCALIBUR, EGI and Swift-HEP

- GridPP + SCD has continued to look for additional sources of funding.
- ExCALIBUR - 6PM to look at Kubernetes to deploy services such as XCaches, Squids, FTS nodes, Rucio Daemons.
  - A second ExCALIBUR funding call will come out in summer.
- EGI funding over the next 2.5 years:
  - 20 PM for Rucio
  - 8 PM for FTS
  - 14.5 PM for CVMFS
- We have 0.5 FTE to work on Rucio from Swift-HEP.

Multi-VO Rucio

- Ian Johnson has been leading Multi-VO Rucio project.
  - Helped by Graduates.
  - New permanent member of staff.
- Aim to move Rucio to a Tier-1 production service 1st half of 2021.
- We will also be building a test data lake with UK Tier-2s over the coming year.



Data Analysis Facility

- The goal of the Vera C. Rubin Observatory project is to conduct the 10-year Legacy Survey of Space and Time (LSST).
- Data taking is due to start in Q4 2022.
  - Data will be sent to a US Institute for initial processing.
  - RAL will be a Data Analysis Centre.
  - IN2P3 will be a major processing centre that will send us data.
- £2million in Capital this year to build a Data Analysis Centre
  - £800k for 9PB of S3 in Echo and 6PB of Tape.
- We have a graduate project funded to start looking at LSSTs needs and hope this will progress to more permanent funding.

Science and
Technology
Facilities Council

Authentication and Authorization

The end of Grid-mapfiles

- The WLCG is pushing forward with efforts to enable the use of tokens across its workflows
  - Starting with data / storage infrastructure
  - Adding CE/pilot workflows
- Many other communities can potentially make use of tokens
  - New communities building workflows for the first time
  - Communities for whom X.509 was not appropriate
- Ensure that the Tier1 is well positioned by the start of Run3 (April 2022)
  - Anticipate co-existence of tokens and X.509 for some time, so plan to that effect

Excerpts from WLCG Token Timeline

Date	Description
May 2021	WLCG baseline services include HTTP-TPC
July 2021	Production IAM Instances Available
Oct 2021	Pilot job submissions may be performed with tokens [Tier1: have something somewhat working]
Dec 2021	VOMS-Admin shutoff. IAM is sole authz provider (including for VOMS server)
Feb 2022	OSG ends support for the Grid Community Toolkit
March 2022	All storage services provide support for tokens
Oct 2022	*Rucio transfers performed with token auth in production*
March 2023	*Experiments stageout & data reads performed via tokens.*
Match 2024	*X.509 client auth becomes optional*

Tasks

- Discovery (April/May)
  - What services use which auth mechanisms (pull out CTA separately)
  - Which (inter)national scale services need to be included in this
    - That we run
    - That are run within the UK (DIRAC)
  - What updated services are needed/available and in what timeframe
  - Identify a good candidate to support this work as a trailblazer
  - What development infrastructure is needed and in what timeframe
- Ongoing HTTP-TPC campaign (May)
- ECHO becoming token capable
  - Engage with wider discussions on access management/etc
  - For xrootd should inherit a solution (engage with early testing)
  - (In general should not need to develop too much on a bespoke basis BUT should engage early and in depth with testing)
  - What are the steps that need to be completed before ECHO -> Production
- CTA deployment

Science and
Technology
Facilities Council

Questions?