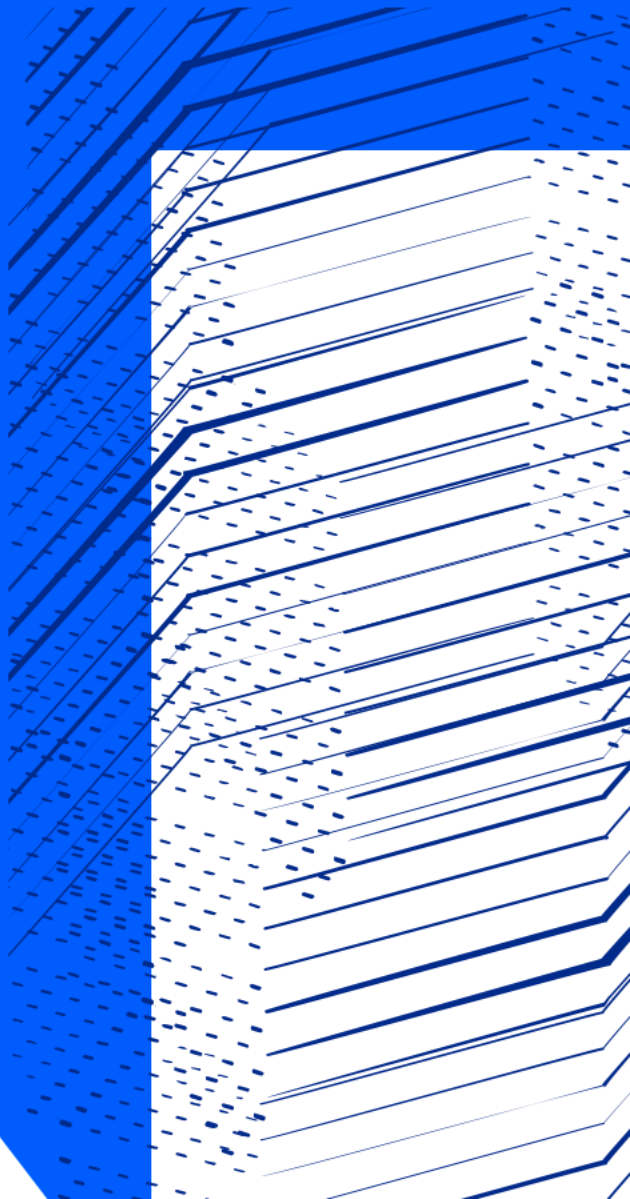
Science and Technology Facilities Council

Tier-1 Plans

Alastair Dewhurst



Introduction

1 Strategy

2 Core Infrastructure

3 Tape

4 Disk

5 CPU

Image © STFC Alan Ford

Science and Technology Facilities Council

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 2026 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.*

How are we doing?

- The Tier-1 has met its MoU commitments.
- We have completed a major upgrade to our network and tape systems.
- Our successful deployment of CTA helped convince FermiLab & DESY to focus on deploying CTA underneath dCache.
- Run joint procurements to reduce costs and consolidate services where possible.
- We have other (not GridPP / IRIS) paying customers for Echo!
- We take advantages of other funding opportunities to jointly run / develop services (Swift-HEP, EGI-ACE, ExCALIBUR).

Risk and Mitigations

- In the GridPP risk register the main risk (I can influence):
  - Failure to retain or recruit key technical staff at RAL
  - Security Incident
  - Failure to procure, deploy or operate hardware at GridPP sites
  - Significant loss of custodial data at the Tier-1
  - GridPP unable to respond to unexpected Technology Shifts
  - Experiment software runs inefficiently, to the detriment of UK physicists
- Setup the PhD conversion Scheme which is currently open.
- David Crooks has setup the SOC (see his talk).
- We ordered hardware early, used direct awards, used new benchmarks.
  - I feel Tier-1 experience could benefit the whole collaboration more.

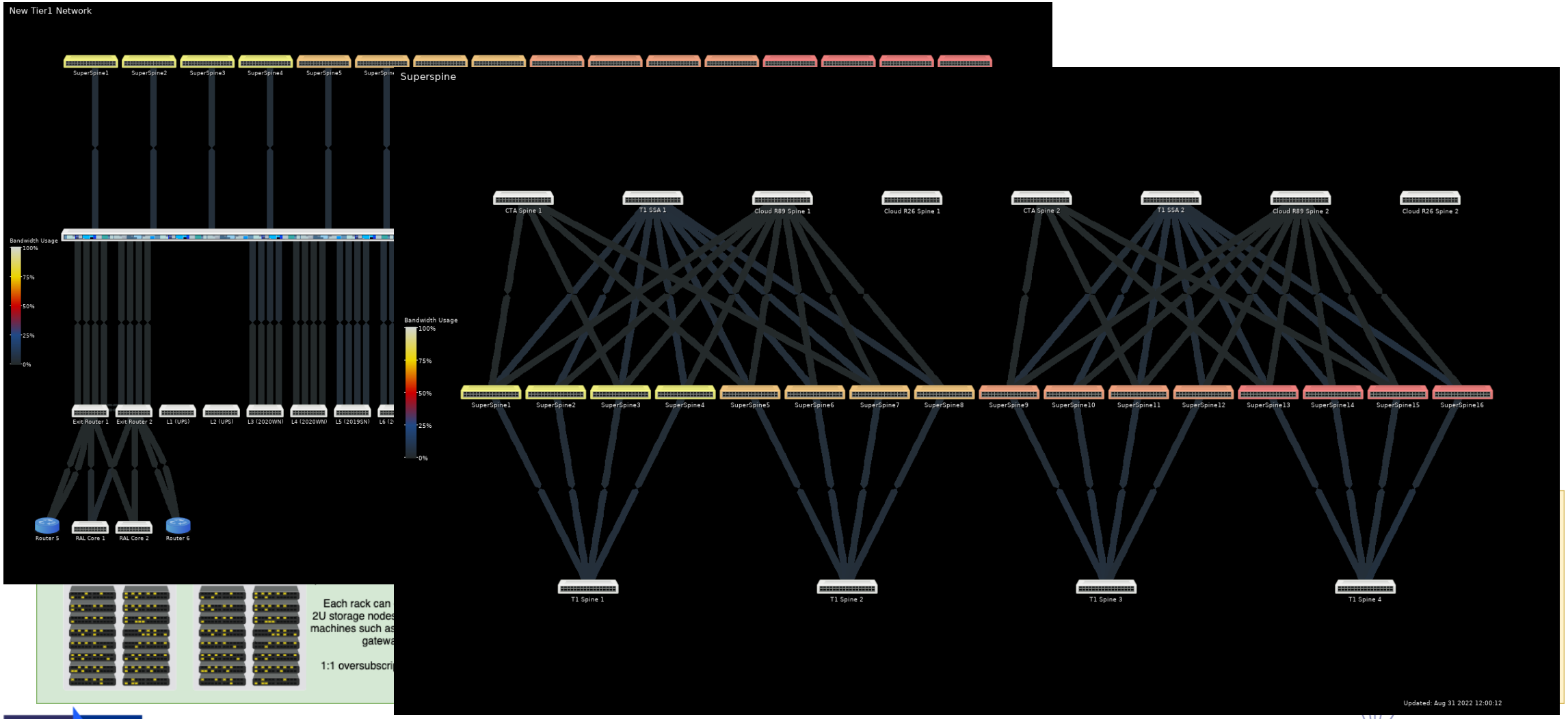
The Research Computing Centre

- The RCC is a new data centre being built at RAL that will be available from early 2026.
- It is designed for high power density computing and will compliment the current R89 data centre.
  - Capacity to support 6.6MW computing load at PUE of 1.1.
  - Aimed for Tier III resilience. 5 min battery backup plus generators.
  - Space for 150 racks with up to 100kW loads per rack.
  - It will have a combination of rear door cooling and direct to chip liquid cooling.
- The RCC is planned to be upgradable to 13.2MW and 19.8MW as demand and budget allows.
- This will be where we want to house our compute in GridPP7.

Network

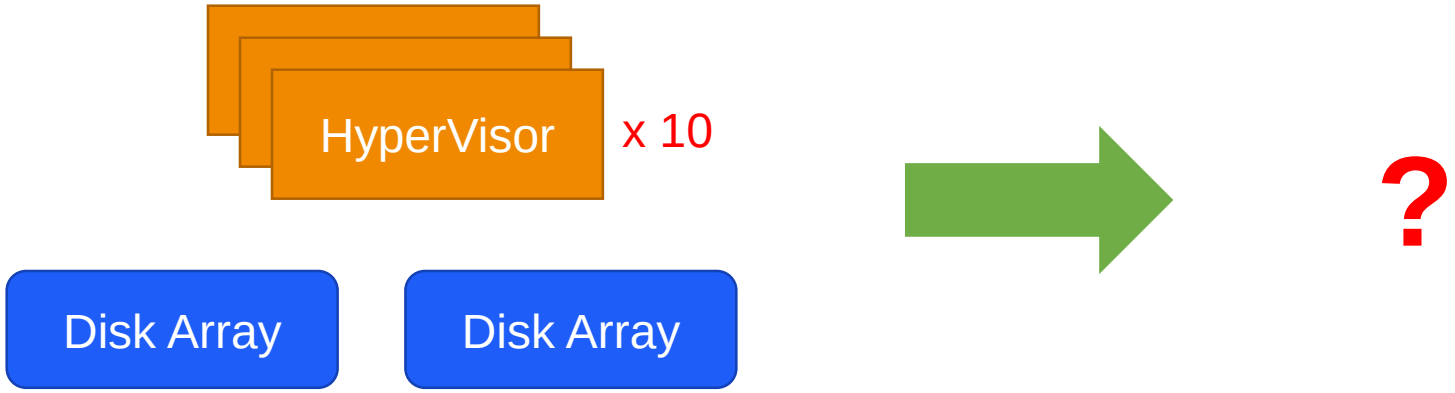
- The Tier-1 has replaced its network.
  - Not all benefits have been realized yet but fundamental limitations in the network are no longer causing operational problems.
  - See James Adam's talk for more details.
- The SCD SuperSpine is now linking projects which allows us to access each other's resources much more easily.
  - We have an SCD network architect.
- We have CHEP paper with a network design out to 2031.
- Still to do:
  - Move batch farm on to LHCONE
  - Get perfSonar boxes working correctly.
  - We will need to upgrade the LHCOPN to 200Gb/s during Run 3.

Network Evolution



Enterprise Virtual Machines

- We have a VMWare Cluster that provides our enterprise VMs that will need to be replaced at the start of GridPP7.
  - ~£300k to replace
- VMWare was recently purchased by Broadcom who are known for increasing support cost prices.
  - Currently we pay <£10k a year for licenses.
- It is not obvious what the best replacement will be:
  - Can it be a shared service with the rest of SCD?
  - What is the most appropriate product / technology to use?
  - Do we need more features, e.g. container orchestration.

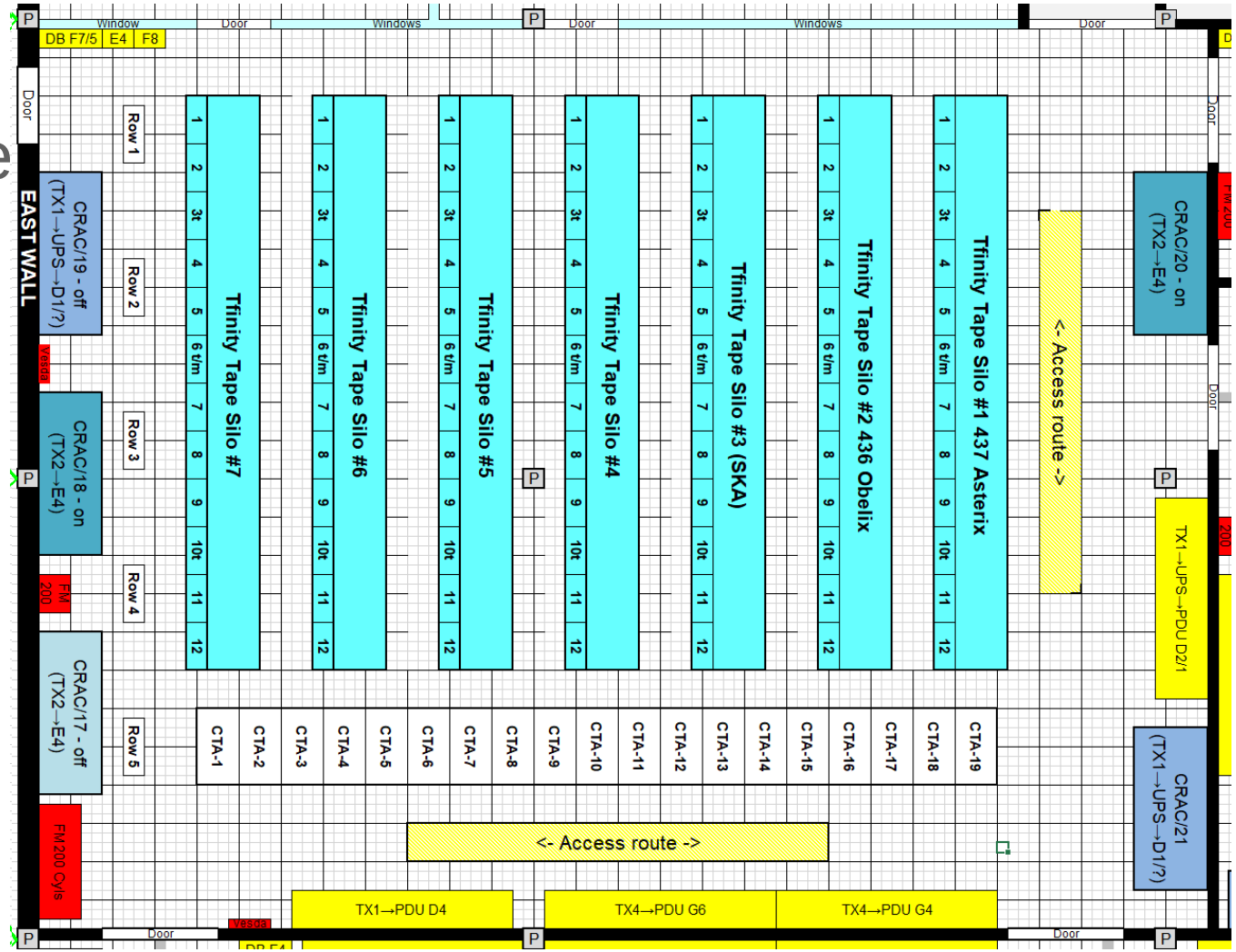


Antares Status and Plans

- Antares was successfully deployed into production in March 2022.
  - All data was successfully migrated for the LHC VOs.
  - We lost a single Tape which held (all of) Minos' data.
- Still to do:
  - Need to properly implement Accounting & Metrics
  - Need to deploy HTTP Tape REST API
  - Need to setup a pre-production instance
  - Need to migrate facilities Castor to Antares
  - Need to migrate from Oracle to PostgreSQL for backend database.

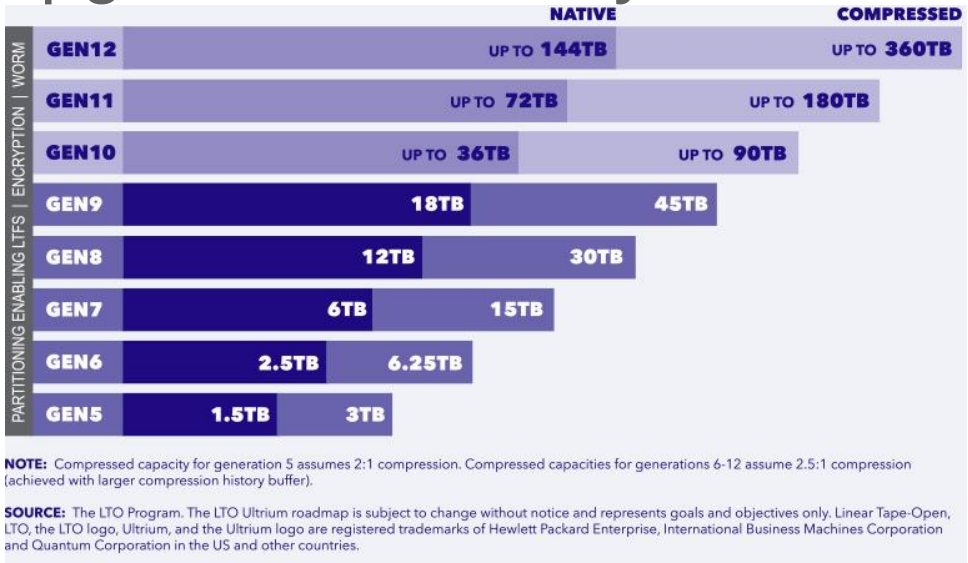
LPD Room Evolution

- The Oracle SL8500 Tape Libraries have a fixed size
- It made sense to put racks at the end of each Library.
- The Spectra Libraries are modular.
- We want to extend them as much as possible.
- Sliding ~~block~~ rack problem.



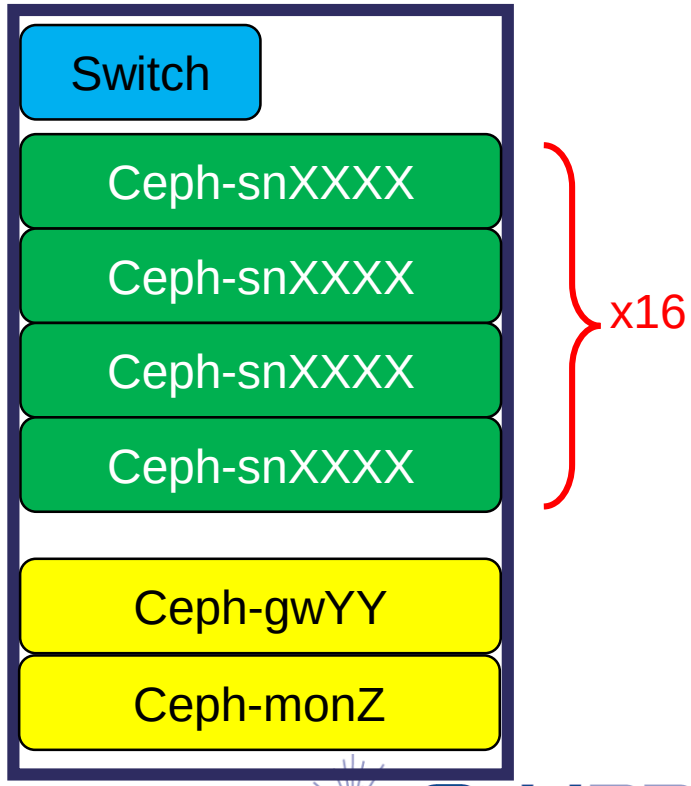
Tape Capacity Plan

- Currently we have a 10 frame Library with a Capacity of 146PB.
  - Expected write by the end of GridPP6 is 143PB.
- We don't want to start GridPP7 with no free capacity.
  - LTO-10 is expected in Q4 2023 and doubles capacity per Tape.
- Before the end of GridPP we want to upgrade our Library to 15 frames and purchase LTO-10 drives.
  - This will effectively double our capacity.
- LTO-11 expected in 2026
- We would start phasing out LTO-9.
- This will increase capacity to 750PB.



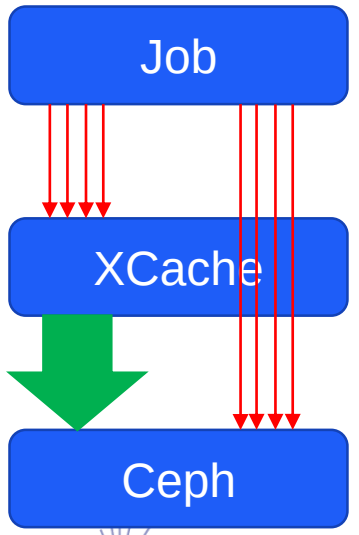
Echo status and plans

- Echo has continued to expand to meet the LHC as well as other VOs capacity requirements.
- We have a very clear architecture and can successfully deploy and decommission large amount of hardware (see Rob's talk).
- Insufficient development effort:
  - Other projects (e.g. Antares) were prioritized.
  - Echo development effort wasn't funded in GridPP6.
  - More to do than anticipated (e.g. retirement of GridFTP).
- Still to do:
  - Upgrade to the latest version of Ceph.
  - Move to Rack Level Failure Domains
  - Automate management of host.



Echo and XRootD

- When Echo was designed, we decided to put Xcaches in-front of Ceph to ‘hide’ the fact it was an object store.
  - This is a sensible concept but hasn’t worked in practise due to them not actually working reliably and introducing many other complications.
- We also hoped that VOs might move to using S3 and while they did move to Webdav for transfers our dependence on XRootD has grown.
- We have migrated to XRootD 5 as well as switched to Webdav for TPC.
- See Jyothish talk for description of technical issues and solutions.

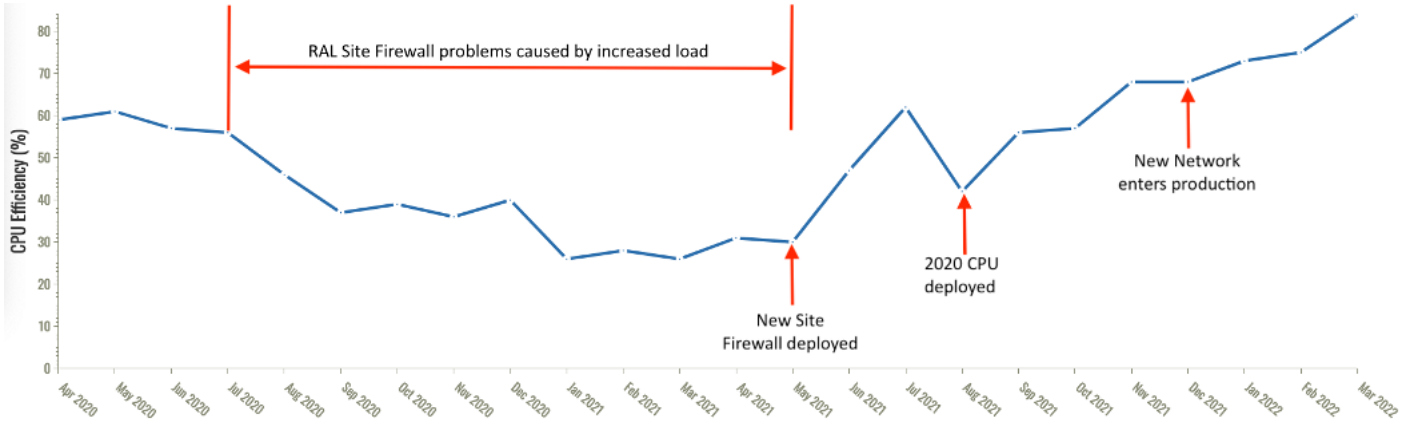


XRootD in GridPP7

- I have spoken to Andreas-Joachim Peters about storage evolution.
  - He thought that continuing with XrdCeph was likely the best course of action.
- RAL will need to maintain a significant amount of expertise (including development) in XRootD.
  - Not just Echo, but for Antares (EOS) and many facilities services that will be built on top of it.
- We need to maintain XrdCeph and LibradosStriper and it seems sensible to combine these.
- We will phase out XCaches in most situations and replace with buffering code.
- We will look to take advantages of benefits of our setup. E.g. parallel transfers.

Batch Farm

- The Batch Farm is currently being upgraded to HTCondor 9 and the WN are being moved to Rocky 8.
  - See a Future Technical Meeting Talk by Tom Birkett!
- Solved a long running CMS CPU Efficiency problem and have SSD storage on the vast majority of the farm.
- Still to do:
  - Move batch farm onto LHCONE
  - Upgrade to ARC CE7 in the next 6 months.



CPU in GridPP7

- CPU is normally the first thing to be reduced if money is tight.
  - I am hopeful that Intel will become competitive with AMD and with supply chains problems easing we will see significant price drops.
- The number of cores per CPU is continuing to grow rapidly and shows no sign of stopping.
  - Higher power density
  - IOPS, networking, memory also need to scale.
- Liquid Cooling may become standard for high end CPUs (and GPUs).
  - Liquid cooling can also increase benchmark results by 2 – 5%
- This year's Tier-1 CPU procurement:
  - Dual AMD EPYC 7763 (64C / 128T each)
  - 1TB memory
  - 480GB SSD for OS and CVMFS Cache
  - 6.4TB NVMe (3DWPD) for job scratch space
  - 25Gb/s NIC.

Summary

- In GridPP6 we have made significant upgrades to our services, which were designed to cope with HL-LHC.
- With delays to HL-LHC we have an excellent foundation for GridPP7.
- We have struggled in areas where we didn't have sufficient development effort.
  - We still need to do those things.

Science and
Technology
Facilities Council

Questions?